EPFL

Thèse n° 8364

Inference methods for the study of interacting biological oscillators in single-cells

Présentée le 25 septembre 2020

à la Faculté des sciences de la vie
Unité du Prof. Naef
Programme doctoral en biologie computationnelle et quantitative

pour l'obtention du grade de Docteur ès Sciences

par

Colas Noé DROIN

Acceptée sur proposition du jury

Prof. P. D. Barth, président du jury
Prof. F. Naef, directeur de thèse
Prof. J. Garcia-Ojalvo, rapporteur
Prof. M. Khammash, rapporteur
Prof. A.-F. Bitbol, rapporteuse

École polytechnique fédérale de Lausanne

2020

À mes parents, pour leur soutien indéfectible.

"Pourtant nous sommes tous pareils... Nous avons quelque chose en commun qui est plus fort que nos différences : c'est le besoin de connaître. Les littérateurs appellent ça l'amour de la science. Moi, j'appelle ça la curiosité. Quand elle est servie par l'intelligence, c'est la plus grande qualité de l'homme."

La Nuit des Temps, René Barjavel

"With equal passion I have sought knowledge. I have wished to understand the hearts of men. I have wished to know why the stars shine. And I have tried to apprehend the Pythagorean power by which number holds sway about the flux."

Autobiography, Bertrand Russel

Acknowledgements

A PhD is always an important step in life, filled with challenges and uncertainty, and I'm extremely thankful for having met people along the way that turned this ordeal into growth and enjoyment. As academic years pass, one comes to realize that science is just as much about exchange and friendship than theoretical readings and alone thinking. I was lucky enough to be surrounded with the right people at the right time, and I owe a lot to the people that supported, taught and guided me, especially at the beginning of these four intense years.

First of all, I'd like to thank my supervisor, Felix, for his guidance, support and openness, despite our different ways of handling theoretical issues. He was almost always available to talk and help when I was stuck with my project, and his advice was deeply relevant. Being myself quite result-driven, I have learnt a lot from his highly careful thinking, in which details matter just as much as the general direction. Felix, without you, the quality of my PhD would have never even got close to what it is now, and I'm grateful for that. Thank you for trusting me during these four years, particularly when I faced difficult times.

Just as importantly, I'd like to thank my colleagues, most of whom I could probably call friends by now. I'll start with my current office mates, Clémence and Nick, as the three of us started almost at the same time and bonded during the first months of our respective contracts. Clémence, your direct and salty humour is always welcome after a day of serious scientific discussion, and I'm happy we could discuss pretty much all aspects of society during our lost hours. Nick, your dry humour hides a big heart, and I'm very thankful for both the help and fun you brought me throughout my PhD. I wish the two of you all the best for the rest of your projects, and I hope you'll stay in Lausanne so that we can share a drink sometimes.

I would then like to thank the people I've been closely working with, namely Eric Paquet, Lorenzo Talamanca and Gioele La Manno. Eric, I've missed your cheerful mood and your bold humour for the last two years, and I wish I words were enough to express how much the delicate touch of your feet under my desk is regretted. Hopefully, I'll see you again someday in Québec. Lorenzo, your stubbornness is almost as great as mine, and our debates were among the best I had. Your wittiness and cooking skills with will be missed just as much as your collaborative spirit and theoretical insights. Gioele, I only got to really know only during the last year of my PhD, but I sincerely thank you for inspiring me with your impressive scientific drive. You're a great person to work with, and your positivity is always welcome when working on challenging projects.

I would also want to give a big thank you to Damien, who ceased to be a colleague a couple of years after I arrived, but who nevertheless remained a friend. Believe it or not, our jogging sessions, as well as our Switch game nights, will probably remain a long time in my memory.

Of course, I cannot forget the Nestle team: Cédric and Ben. Cédric, you've been a great office-mate, and the quality of your foot touch was almost as great as Eric's. Your inexhaustible humour is precious, thank you for easing the rough months I was going through when we were sharing an office. Ben, for the sake of authenticity, I can't really compliment your german humour, but I can compliment your perseverance in trying to make us laugh with it (and this, in itself, was funny). You're a good man, say hello to the kangaroos for me. All the best to you two.

I also want to thank all my other colleagues, which I less had the chance to get to know, but who still participated to the great lab atmosphere: Hugo, Irene, Nagammal, Alex L, Eric D, Onur, Jerome, Jonathan S, Laura, Daniel, Jake, Jingkui and Ambroise. Also, a big thank to Sophie, who made my administrative life much easier and was always helpful when I needed her. Besides, I'd like to thank all the interns: Dan, François, Alice and Alexis. Similarly, I can't forget the (amazing) TA team: Daniel, Olivia, Marine, Giovanna, plus all those that I have thanked already.

I would then like to thank the friends I got to know in Lausanne, which made my stay enjoyable: Romain, Sylvain, Thomas, Silvia, Madhi, Kato, Mahé, Elias, Alexandra W, Aleksandra M, plus all these people that I'm forgetting forget but with whom I regularly chit chat in the corridors. I hope I'll keep seeing all of you after my PhD.

Je voudrais aussi citer les amis que j'ai connus plus jeune, dont la plupart sont venus me voir à l'EPFL : Hassan, Alice, Xavier, Clément J, Clément D, Alexandre, Morgane, Léon, et bien d'autres encore. Merci à tous pour le fun et la diversité que vous avez apportés à ma vie.

Il m'importe tout autant de remercier les membres de ma famille, au sens large du terme, pour la gentillesse, le soutien et le fun dont ils ont fait preuve : mes cousin(e)s, oncles et tantes. Mamila, pour son amour intarissable. De même, mes grand-parents décédés, qui, malheureusement, ne me verront jamais devenir docteur. Un merci sincère et chaleureux, pour les réunions de familles, les voyages, les débats passionnées et les blagues inoubliables.

Finallement, par dessus tout, je voudrais remercier ma famille proche. Mon frère et ma soeur, Jonathan et Léa, et mes parents, Annick et Philippe, pour leur amour et leur soutien inconditionnels. Ils savent, plus que quiconque, les problèmes de santé que j'ai traversés, et ont toujours essayé de m'aider. Vous m'avez appris ce qu'il y a d'important. Vous avez fait de moi qui je suis. Un millier de fois, merci.

Lausanne, June 17th, 2020

Colas

Abstract

Biological oscillators are pervasive in biology, covering all aspects of life from enzyme kinetics reactions to population dynamics. Although their behaviour has been intensively studied in the last decades, the recent advances of high-throughput experimental technologies in the fields of omics and microscopy has called for the development of new analysis methods. Among the many types of models and quantitative analyses, parametric approaches are promising as they enable for a mechanistic or physical explanation of the phenomena under study. In particular, dynamical systems theory seems particularly adapted as the vast majority of oscillators can be modelled through differential equations. Dynamical systems parameters can also be easily optimized *via* maximum likelihood approaches. The validity of the inferred model can then be assessed from the quality of its predictions. We here present three different scientific questions regarding noisy biological oscillators, which are answered using maximum-likelihood inference approaches applied to parametric models.

We first take interest in the characterisation of the influence of the cell-cycle over the circadian clock in individual mammalian cells. To this end, we develop a method combining a Hidden Markov Model with an Expectation-Maximization algorithm to infer their coupling from single-cell microscopy traces. We show that this coupling predicts multiple phase-locked states exhibiting different degrees of robustness against molecular fluctuations inherent to cellular scale biological oscillators.

We then try to understand how the mammalian transcriptome behaves in the liver. Thence, we use single-cell RNA sequencing (scRNA-seq) along with mixed-models to investigate the interplay between gene regulation in space and time. Categorising mRNA expression profiles using mixed-effect models and smFISH validations, we find that many genes in the liver are both zonated and rhythmic, most of them showing multiplicative space-time effects.

Finally, we look more closely at the cell-cycle, as it is one of the main drivers of gene expression cell-to-cell heterogeneity in otherwise homogeneous cell populations. Here, we would like to understand if and how cell-cycle velocity changes depending on the phase of the cycling cells. To that end, we formulate the problem in terms of an autonomous dynamical system and use this to infer consistent dynamics for the cell-cycle from scRNA-seq data.

Phase inference being paramount in all of these three studies, a short technical review on the topic is also provided at the end of this thesis, along with Julia scripts for the main inference

methods presented. Various computational tools assisting the understanding of the scientific questions at stake are also presented, including a Python Dash web-app, a D3 widget and many Matplotlib animations and widgets.

Keywords

biological oscillators, inference, optimization, maximum likelihood, dynamical systems, nonlinear physics, modelling, quantitative biology, single-cell biology, hidden Markov model, HMM, expectation-maximization, EM, coupled oscillators, stochastic dynamics, phase-locking, synchronization, circadian clock, cell-cycle, liver, single-cell RNA sequencing, scRNA-seq, gene regulation, gene expression, mixed-effect model, model selection, Akaike Information Criterium, AIC, single-cell transcriptomics, RNA velocity, dimensionality reduction, cell-cycle dynamics, phase inference, Python, Python Dash, Julia, Matplotlib

Résumé

Les oscillateurs biologiques sont omniprésents en biologie, couvrant tous les aspects de la vie, des réactions cinétiques enzymatiques à la dynamique des populations. Bien que leur comportement ait été intensivement étudié au cours des dernières décennies, les récentes avancées des technologies expérimentales à haut débit dans les domaines de l'omique et de la microscopie a récemment appelé au développement de nouvelles méthodes d'analyse. Parmi les nombreux types de modèles et d'analyses quantitatives, les approches paramétriques sont prometteuses car elles permettent une explication mécanistique ou physique des phénomènes étudiés. En particulier, la théorie des systèmes dynamiques semble particulièrement adaptée car la grande majorité des oscillateurs peut être modélisée au travers d'équations différentielles. De plus, les paramètres de tels systèmes peuvent être facilement optimisés par des approches de type maximum de vraisemblance. La validité du modèle inferré peut alors être évaluée à partir de la qualité de ses prédictions. On présente ci-après trois questions scientifiques concernant des oscillateurs biologiques bruités, que l'on traite par des approches d'inférence de type maximum de vraisemblance appliquée à des modèles paramétriques.

On s'intéresse d'abord à la caractérisation de l'influence du cycle cellulaire sur l'horloge circadienne dans des cellules uniques de mammifères. Ainsi, on développe une méthode combinant un modèle de Markov caché et un algorithme de type espérance-maximisation pour inférer leur couplage à partir de traces cellulaires provenant de microscopie. On montre que ce couplage prédit plusieurs états de verrouillage de phase, présentant différents degrés de robustesse vis-à-vis des fluctuations moléculaires inhérentes aux oscillateurs biologiques à l'échelle cellulaire.

On essaye ensuite de comprendre comment le transcriptome des mammifères se comporte dans le foie. Pour cela, on utilise le séquençage d'ARN de cellules uniques (scRNA-seq) en combinaison avec des modèles à effets mixtes pour étudier l'interaction entre la régulation des gènes dans l'espace et le temps. En catégorisant les profils d'expression d'ARNm à l'aide de modèles à effets mixtes et de validations smFISH, on constate que de nombreux gènes du foie sont à la fois zonés et rythmiques, la plupart d'entre eux exhibant des effets multiplicatifs spatiotemporels.

Enfin, on s'intèresse de plus près au cycle cellulaire, puisque c'est l'un des principaux facteurs d'hétérogénéité entre cellules de l'expression des gènes dans des populations cellulaires par ailleurs homogènes. Dans ce cadre, on voudrait comprendre si et comment la vitesse du cycle

cellulaire change en fonction de la phase des cellules qui cyclent. Pour cela, on formule le problème en termes de système dynamique autonome et l'utilise pour déduire une dynamique cohérente à partir de données de scRNA-seq.

L'inférence de phase étant primordiale dans ces trois études, une brève revue technique sur le sujet est également fournie à la fin de cette thèse. Divers outils informatiques permettant de comprendre les questions scientifiques en jeu sont également présentés, notamment une application web de type Python Dash, un widget D3 et de nombreuses animations et widgets faits avec Matplotlib.

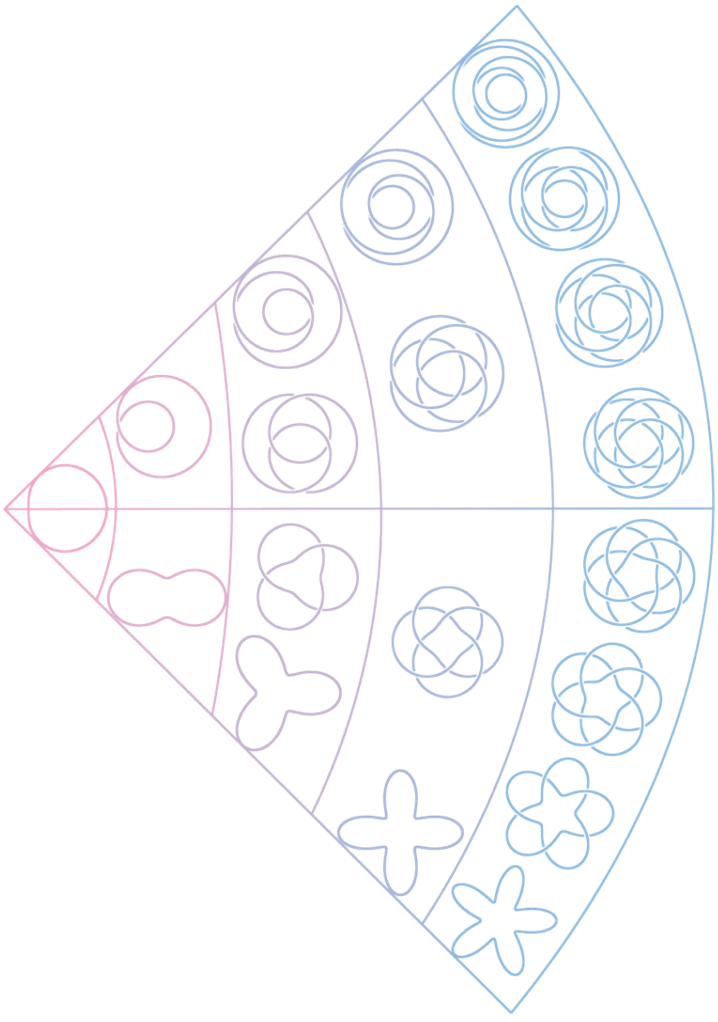
Mots-clés

oscillateurs biologiques, inference, optimisation, maximum de vraisemblance, systems dynamiques, physique des systems non-linéaires, modélisation, biologie quantitative, biologie des cellules uniques, modèle de Markov caché, HMM, espérance-maximisation, EM, oscillateurs couplés, dynamique stochastique, verrouillage de phase, synchronisation, horloge circadienne, cycle cellulaire, foie, séquençage de l'ARN des cellules uniques, scRNA-seq, régulation génétique, expression génétique, modèle à effets mixtes, sélection de modèles, Critère d'Information d'Akaike, AIC, transcriptomique des cellules uniques, vélocité de l'ARN, réduction de la dimensionnalité, dynamique du cycle cellulaire, inférence de phase, Python, Python Dash, Julia, Matplotlib

Table of Contents

Introduction

This introduction is an original work, summarizing some of the knowledge accumulated during my PhD in the Naef Lab. Nevertheless, its writing was also inspired by several textbooks, which, out of respect for the authors, I wish to mention. This includes *Synchronization: A Universal Concept in Nonlinear Sciences* by Arcady Pikovsky[1], *Nonlinear Dynamics And Chaos* by Steven Strogatz[2], *Biological Clocks, Rhythms, and Oscillations* by Daniel Forger[3], *An Introduction to Systems Biology: Design Principles of Biological Circuits* by Uri Alon[4], and finally *Biological Timekeeping: Clocks, Rhythms and Behaviour* by Vinod Kumar[5].



Artwork Figure 1: Artistic representation of the phase-lockings observed in a system of coupled oscillator showing stable oscillations. The kind of phase-locking representation is known as $p:q$ *torus knots*, that is, torus trajectories viewed from above and projected on a plane. From left to right, bottom to top, represented trajectories are: 1:1, 1:2, 2:1, 3:1, 3:2, 2:3, 3:2, 1:4, 3:4, 4:3, 4:1, 1:5, 2:5, 3:5, 4:5, 5:4, 5:3, 5:2, 5:1.

Outline

In Section 1, I briefly introduce biological oscillators from a contextual point of view, along with some of the very basic mathematical concepts used to describe oscillation properties (phase, amplitude, period).

In Section 2, I introduce, with intuitive terms, the concepts of coupling, synchronization and phase-response curve for simple phase oscillators, along with some example taken from biological systems.

In Section 3, I introduce the more advanced mathematical framework used to describe oscillator and synchronization from a dynamical systems perspective: conditions needed for oscillations, linear stability analysis, Hopf bifurcation, Poincaré map, Arnold tongues.

In Section 4, I explain how theoretical biological models can be fitted to experimental data, what knowledge can be gained from this fitting, and what are the corresponding issues: likelihood, overfitting, model transparency and model tractability.

In Section 5, I introduce the four corresponding applications presented in the rest of this thesis: the modelling and inference of the influence of the cell-cycle on the circadian clock, the classification of the liver transcriptome behaviour into different spatio-temporal categories, the inference of cell-cycle velocity from static datasets and a technical review on phase inference methods.

1. What is a biological oscillator?

1.1. General introduction

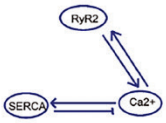
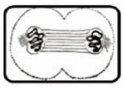
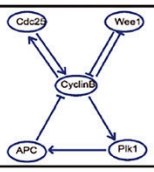
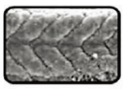
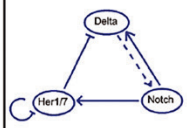
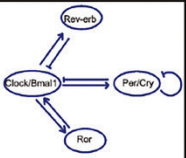
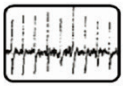
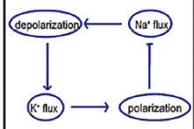
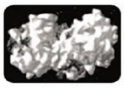
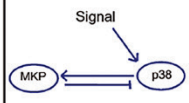
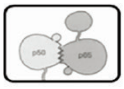
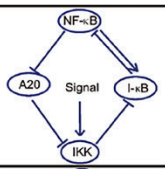
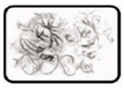
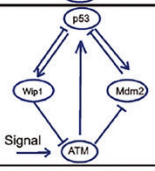
The topic of biological oscillators is vast, covering aspects of life ranging from simple enzyme kinetics reactions to population dynamics. This includes complexes processes as varied as calcium signalling, cell division, heart rate and breathing, or even periodic outbreaks of diseases[6]–[10]. Overall, generated rhythms are essential to most organisms, from cyanobacteria to mammals and from biochemistry to mood regulation[11]–[13].

The scope of this thesis will be limited to the study of rhythms occurring within biological organisms. Still, it should be noted that oscillators properties do not depend much on the context. Therefore, the methods we will present here could also apply to the study of nonbiological rhythms. Overall, it is only the system that changes, but not the emerging oscillating behaviour. The mechanisms of rhythm generation in nonbiological applications are extensively described in the literature[14], [15]. Nonetheless, it remains true that biological rhythms themselves often depend on nonbiological factors (for instance, solar photic cues). Sometimes, things can be trickier as in the case of the approximately 24-hour periodic emergence of fruit flies from their pupae, which might appear to be governed by the external daily rhythm, but actually stems from an endogenous clock[16].

Still, endogenous rhythms also cover a broad range of biological systems, with periods ranging from milliseconds (e.g. neuronal spikes[17], vocal folds oscillations[18]) to years (hibernation cycles[19]). Such rhythms can be generated at very different levels: physiological (heart pacemaker, respiration, hormones)[20], in intracellular biochemical networks (calcium oscillations, glycolytic oscillations)[21], or via transcriptional feedback loops (somitic clock, NF-kB oscillations, circadian clock)[22], [23]. Introduction Figure 1 gives a good overview of the variety of rhythms observed at the microscopic scale.

Oscillators having a 24h periodicity are especially important. These clocks, which are called circadian (*circa*, around, *dian*, a day), are fundamental to life and are present in almost all complex organisms[1]. They are introduced in details in Chapter 1, Section 1.2.1.1.

[1] Viruses and bacteria (except for cyanobacteria) do not have a clock.

Oscillators	Network descriptions	Period	Oscillator type	Function	Ref.
 Cardiomyocytes calcium spike		~1 s	Membrane oscillators	Endogenous pacemaker	Periasamy *et al.* 2008 [3]; Liu *et al.* 2008 [4]
 *Xenopus* embryo cell cycle		10 min−24 h	Cytoplasmic oscillators	Endogenous pacemaker	Yang *et al.* 2013 [1]
 Zebrafish segmentation clock		25 min (zebrafish) 90 min (chicks) 2 h (mice)	Genetic oscillators	Endogenous pacemaker	Mara and Holley 2007 [6]
 Mammal circadian clock		24 h	Genetic oscillators	Endogenous pacemaker	Bell-Pedersen *et al.* 2005 [7]
 Action potential		0.001 s−10 s	Membrane oscillators	Signal processing and transduction	Hodgkin and Huxley 1952 [8]
 p38 oscillation		~60 min	Cytoplasmic oscillators	Signal processing and transduction	Tomida *et al.* 2015 [9]
 NF-κB spike		~100 min	Genetic oscillators	Signal processing and transduction	Zambrano *et al.* 2016 [10]
 p53 oscillation		~6 h	Genetic oscillators	Signal processing and transduction	Batchelor *et al.* 2011 [11]

Introduction Figure 1: Summary of core architectures, periods, types and functions observed in oscillators originating from various biological systems. Taken from *Li and Yang, 2018* [24].

1.2. History

The first historical trace of biological oscillator science probably goes back to the Swedish naturalist Carolus Linnaeus, who decided to use the observed plant circadian rhythms to create a garden whose flowers would open and close at different times of the day. To this end, Linnaeus recorded, over many years, the daily rhythms of more than thirty

plant flowers. In 1751, he finally designed his *Horologium Florae* (Introduction Figure 2), which, unfortunately, was never planted while he was alive.

Introduction Figure 2: Representation of the *living flower clock* imagined by the Swedish naturalist Carolus Linnaeus, adapted from his book *Philosophia Botanica* (1751).

In parallel, in France, Jean-Jacques Dortous de Mairan started to experiment on haricot beans and mimosas. In 1729, he noticed that the leaves of the plant moved up and down depending on the time on the day, and deduced from that that an internal clock had to be at work inside of the plants. But the analytical study of oscillations really dates from the mathematician Alfred James Lotka, who put forward a theoretical reaction which exhibits damped oscillations (1910). About a decade later, he proposed the reaction mechanism which now carries the Lotka–Volterra label. In parallel, oscillations were observed by Bray (1921) in the hydrogen peroxide–iodate ion reaction. Since that time, much more complex experiments have been performed in different laboratories[1]. In the 1950s, the first clear examples of biochemical oscillations were recognized in glycolysis, in cyclic AMP production, and in the horseradish peroxidase reaction[25]. Another significant discovery of an oscillating reaction was made by Belousov in the late fifties, the study of which was continued by Zhabotinskii in 1969 and is now known as the Belousov–Zhabotinskii reaction[26]. In the 1980s and 1990s, with the great progress made in molecular biology, many proteins and genes

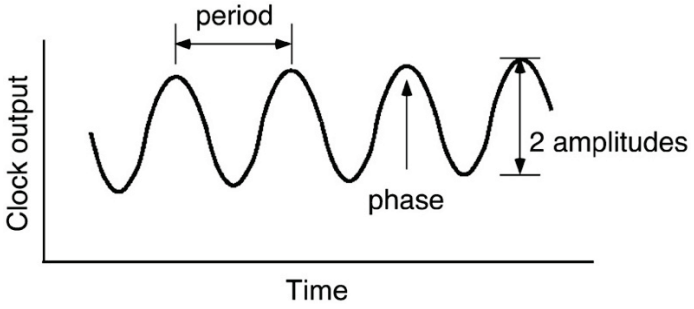
interaction networks generating oscillations were discovered, such as the PERIOD and CRY proteins in the mammalian circadian clock[27], and the Cyclin proteins in the eukaryotic cell-cycle control[28].

1.3. Basics: oscillators, oscillations, clocks and rhythms

Before providing an in-depth explanation of the mathematical concepts used to describe oscillators (cf. Section 3), I here introduce the basics using more intuitive definitions.

The meaning of oscillators, oscillations, clocks and rhythms can vary depending on the context. One example among many is that a *clock* is usually understood as insensitive to temperature in biology[2] (because it keeps track of time, and therefore should be robust to physiological or environmental fluctuations), while in physics or engineering this requirement does not usually exist. In the rest of this work, we'll use the word *clock* only to refer to the circadian clock (which *is* temperature compensated[29]), and *oscillators* to refer to the more general class of all rhythmic systems (which, of course, include the circadian clock). *Rhythms* and *oscillations* refer to the outputs of oscillators.

The *Period* refers to the inherent time scale of an oscillator, measuring the time needed for the oscillator to complete one full oscillation. If the system period is stable over time oscillations can be used to measure time. In practice, only very few systems are robust enough to provide a reasonable time estimate. For instance, the human circadian clock tends to have a period that not only vary among individuals, but which also vary in time for a given individual[30]. Introduction Figure 1 provides an overview of the very vast range of periods observed in biological systems. Introduction Figure 3 illustrates the notion of period with a simple cartoon.



Introduction Figure 3: Critical terms to describe oscillations: period, phase and amplitude. Taken from reference [31].

It is essential to distinguish the natural or intrinsic period of an oscillator from its period when coupled with an external system. Indeed, the behaviour of an oscillator can significantly

[2] This it not true for the segmentation clock, but it's not a clock *stricto sensu*.

vary depending on how it interacts with its environment (cf. Section 2), and its natural period (the one it has when ticking in isolation) can be very different from its period after coupling.

When studying an oscillating system, one is often interested in quantifying the fraction of a cycle that has elapsed between two events (e.g. peaks in a time series). It's precisely the purpose of the *phase*, which, within a cycle, measures how much progress the oscillator has made. Therefore, the phase can always be simplified modulo the oscillator period. Phase units can vary depending on the problem, but the concept is always the same whether it is provided in radian (progression along the polar circle, with length 2π), in time units (progression along the period T), or even in percent. Introduction Figure 3 illustrates the notion of phase with a simple cartoon. As a rule, a given phase always relates to a given state of the corresponding oscillatory system. Therefore, the phase can also be considered as a 1-dimensional coordinate used to parametrize the state of an oscillating system. However, in dynamical systems, the mapping between the high-dimensional system to the one-dimensional phase can be quite complex, such that the interpretation of the concept of phase is usually not obvious. This is developed in more details Sections 2.3 and 3.1.

One last characteristic of oscillations is *amplitude*. If the oscillating system is conservative[3], its amplitude will be uniquely determined by its internal energy (cf Section 3.2). Introduction Figure 3 illustrates the notion of amplitude with a simple cartoon[4].

2. Coupling and synchronization

2.1. General introduction

Biological systems always interact with each other and with the external environment (providing an energy source). In the case of cycling systems, one speaks of *synchronization* when their interaction leads to concomitant oscillations, although the exact definition is more complicated.

Overall, synchronization is a broad topic and have been the object of numerous papers[32]. Biological oscillators are particularly interesting in this regard: heartbeat is partly synchronized with respiration and movement[33], circadian rhythms are synchronized with light

[3] This almost never happens in biology, as living organisms are always in interaction with their environment and depend on external sources of energy.

[4] Note that there's a subtility between height and amplitude, as the latter is defined as twice the vertical peak to trough distance.

cycles and movement (among other things)[34], neurons fire in order as signals travel throughout the brain[35]. Synchronization can also be observed between different organisms: women living together appear to start their menstruations at the same time[36], and same goes for the harmonious blinking of fireflies[37].

Conversely, synchronization disruption is very often associated with diseases. Arrhythmia can lead to severe cardiovascular problems[38], nocturnal work (desynchronizing the circadian clock) is related to an increase in breast cancer[39], epilepsy can lead to sudden death[40], etc.

I here introduced the concepts of coupling and synchronization using mostly intuitive definitions. For a more mathematical approach of the subject, refer to Section 3.4.

2.2. History

The etymology of the word *synchronization* can be traced back from the Greek $\sigma\upsilon\gamma\chi\rho\upsilon\acute{\iota}\zeta\omega$, which literally means "the same time". The scientific history of synchronization goes back to as far as the 17th century, when the Dutch mathematician and physicist Christian Huygeens started experimenting on the behaviour of two coupled pendulum clocks. He acknowledged that two pendula set next to each other, on the same unstable support, tended to become synchronized, swinging in opposite directions. Consequently, he assumed that the support was naturally transmitting information between the pendula, therefore acting as a coupling and enabling synchronization. A few centuries later, in the middle of the 19th century, John William Strutt Rayleigh described the synchronization, but also the quenching (i.e. the death of oscillations), of organ pipes in his *Theory of Sound*. When electrical and radio engineering started emerging, the investigation of synchronization suddenly accelerated. In 1920, W. H. Eccles and J. H. Vincent applied for a patent confirming their discovery of the synchronization property of a triode generator. During the 1920s, the first theoretical study of the synchronization was made by Appleton and Van der Pol. After this, the literature on synchronization started becoming denser, and it gradually became clear that diverse phenomena, which didn't seem to have anything in common (e.g. the sound of organ pipes and the songs of snowy tree crickets) obeyed the same universal laws[1].

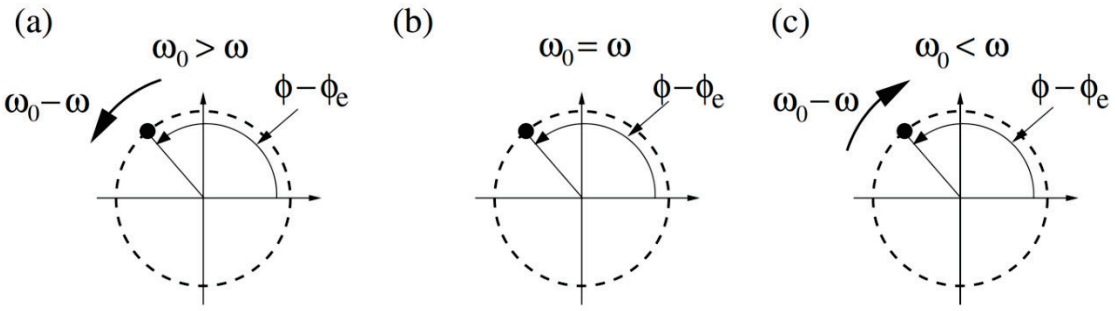
2.3. What is coupling?

Synchronization is not possible if oscillators are non-interacting. This interaction, between two or many oscillators, is called coupling and can come in many shapes. One way of describing a coupling depends on the relationship between the oscillators of the system under study: if the system is made of two oscillators, either the coupling is unidirectional, either bidirectional; if more oscillators interact, more complex interaction schemes are usually

considered. Another essential feature of coupling is when it is active. If it is continually sending a signal to other oscillator(s), it will be termed *continuous* (e.g. two parts of the same biochemical network). Conversely, when it occurs only at a specific phase (or combination of phases), it is termed *pulsed* (for instance, the light from the sun only impacts the circadian clock during the day).

2.3.1. Unidirectional coupling

In the case of a unidirectional coupling, represented Introduction Figure 4 below, one oscillator with phase ϕ is coupled to a periodic external action with phase ϕ_e. In a simple setting, the coupling can be represented by a force that tends to drive the speed ω of the oscillator towards the entraining speed ω_0. Three cases are then possible: either the coupling is not strong enough and no synchronization is observed, meaning that ω_0 is superior (a) or inferior (c) to ω, leading to a decreasing or increasing (respectively) phase difference between the two; either the coupling can synchronize the oscillator to the external action (b), leading to an identical average frequency ($\omega_0 = \omega$) and a phase difference that doesn't tend to grow in time[5]. In case (a), the coupling is said to be positive or *excitatory*, meaning that it tends to accelerate the receiving oscillator. In case (c), the coupling is said to be negative or *inhibitory*.



Introduction Figure 4: Representation of the three different possible cases observed in case of a single oscillator under the influence of an external coupling. Either the external coupling with frequency ω_0 tend to accelerate (**a**) or decelerate (**c**) the natural speed ω of the oscillator, and the phase difference $\phi - \phi_e$ decreases or increases in time (respectively), either the coupling leads to synchronization (**b**), which means that the frequencies get identical, and the phase difference is constant in time. Visual taken from the book by *Pikovsky*[1].

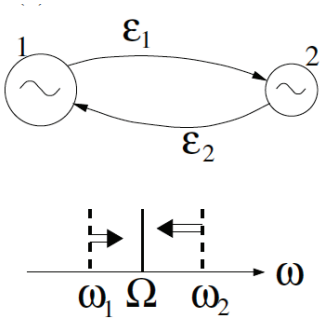
Other settings are possible, since a coupling can be both excitatory and inhibitory depending on the phase of the two systems (cf. Figure 1.2, from Chapter 1), and can lead to complex ratios (or quasiperiodicity) between the periods of the giving and receiving system (see

[5] In practice, it can vary along one cycle, but it should be constant, cycle after cycle, when observed at a given time of the cycle.

Section 3.4). Besides, the relationship between phase advance/delay and coupling type is not trivial, as excitatory couplings can lead to phase delay at steady-state, and conversely. This is due to the circular nature of the phase space and will depend on both the strength of the coupling and of *when* it is applied. In all of these cases, since the coupling is unidirectional, one uses the term *entrainment* rather than synchronization.

2.3.2. Bidirectional coupling

Bidirectional coupling is richer as oscillators influence each other mutually and this can lead to a faster and more robust synchronization, along with more complex phase behaviours. In the most straightforward setting, the coupling will make the average frequencies of the two oscillators, ω_1 and ω_2 closer to each other, until synchronization happens[6], in which case $\omega_1 = \omega_2 = \Omega$ (Introduction Figure 5). Note, however, that Ω is not necessarily the average of the two intrinsic frequencies, as its value will depend on the strength and time-dependence of the couplings. In more complex settings, complex ratios or quasiperiodity can be observed between ω_1 and ω_2. This is developed in the next section.



Introduction Figure 5: Representation of the mutual coupling of two oscillators. When interacting, oscillators 1 and 2 will tend to synchronize their average frequency ω_1 and ω_2. Synchronization is reached when $\omega_1 = \omega_2 = \Omega$ over any time window. Visual taken from the book by *Pikovsky*[1].

2.3.3. Complex interaction

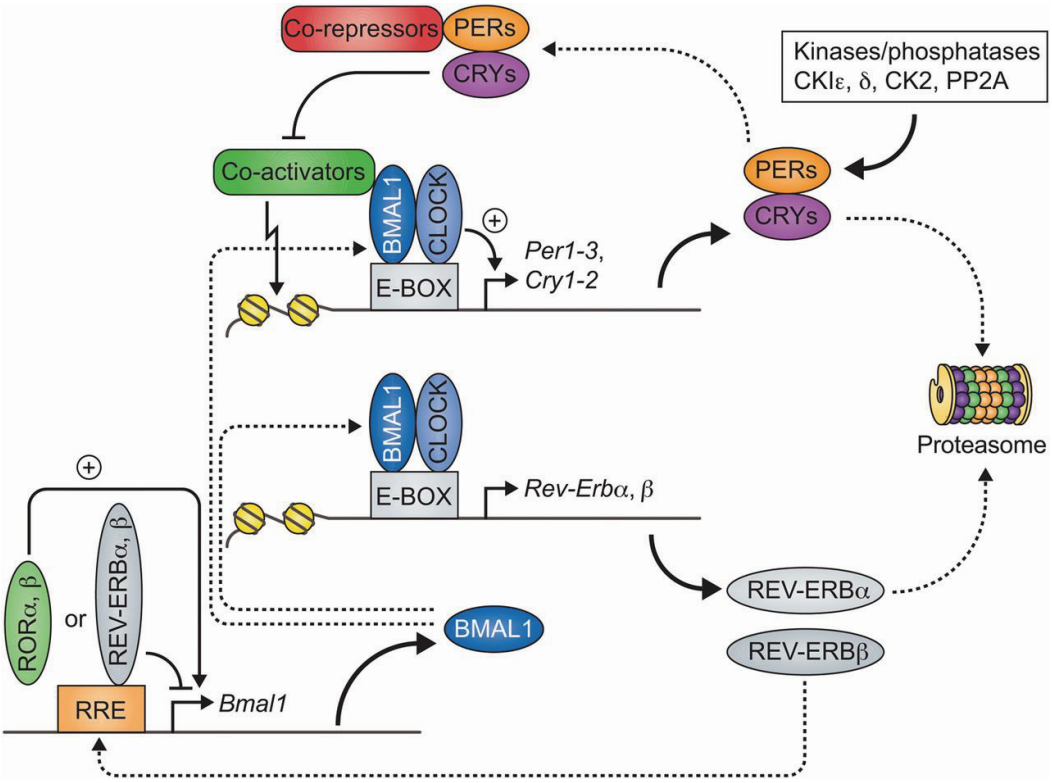
When more than two oscillators are considered, complex interaction schemes can exist. Among them, noteworthy ones are mean-field (each oscillator affects all other oscillators), nearest-neighbour (each oscillator only affects its neighbours), random coupling (oscillators affect each other with a given probability), small-world network coupling (oscillators affect each other in small groups (cliques), but these groups also affect each other but with

[6] It's important here to distinguish the instantaneous frequencies (which vary in time, and are not equal most of the time) from the average frequencies, which *are* equal in the situation described above. The average frequencies are obtained by averaging the instantaneous frequencies over many cycles (in theory, an infinity).

less coupling strength/probability)[41], and finally scale-free networks coupling (the probability of being connected to other oscillators follows a power-law).

2.3.4. Identifying the relevant degrees of freedom

In some systems, each oscillator can easily be identified, and the coupling can be singled out as well. This is the case in most brain simulations, where each neuron is physically distinct from its neighbours, and all the communication between them occurs through synapses[42]. However, in many other systems, things are much harder to disentangle. For instance, in biochemical networks, two feedback loops can share a common element (cf Introduction Figure 6 below). In this case, it would be true to say that the whole network is one unique oscillator. Still, in many cases, it would also be relevant to consider the subsystems as separate, coupled oscillators, each with its own phase, amplitude and period.



Introduction Figure 6: Molecular representation of the core circadian clock, in which several feedback loops (PER-CRY, REV-ERB, ROR) share commons elements (BMAL, CLOCK, RRE). Taken from *Brown et al., 2014* [43].

2.4. What is synchronization?

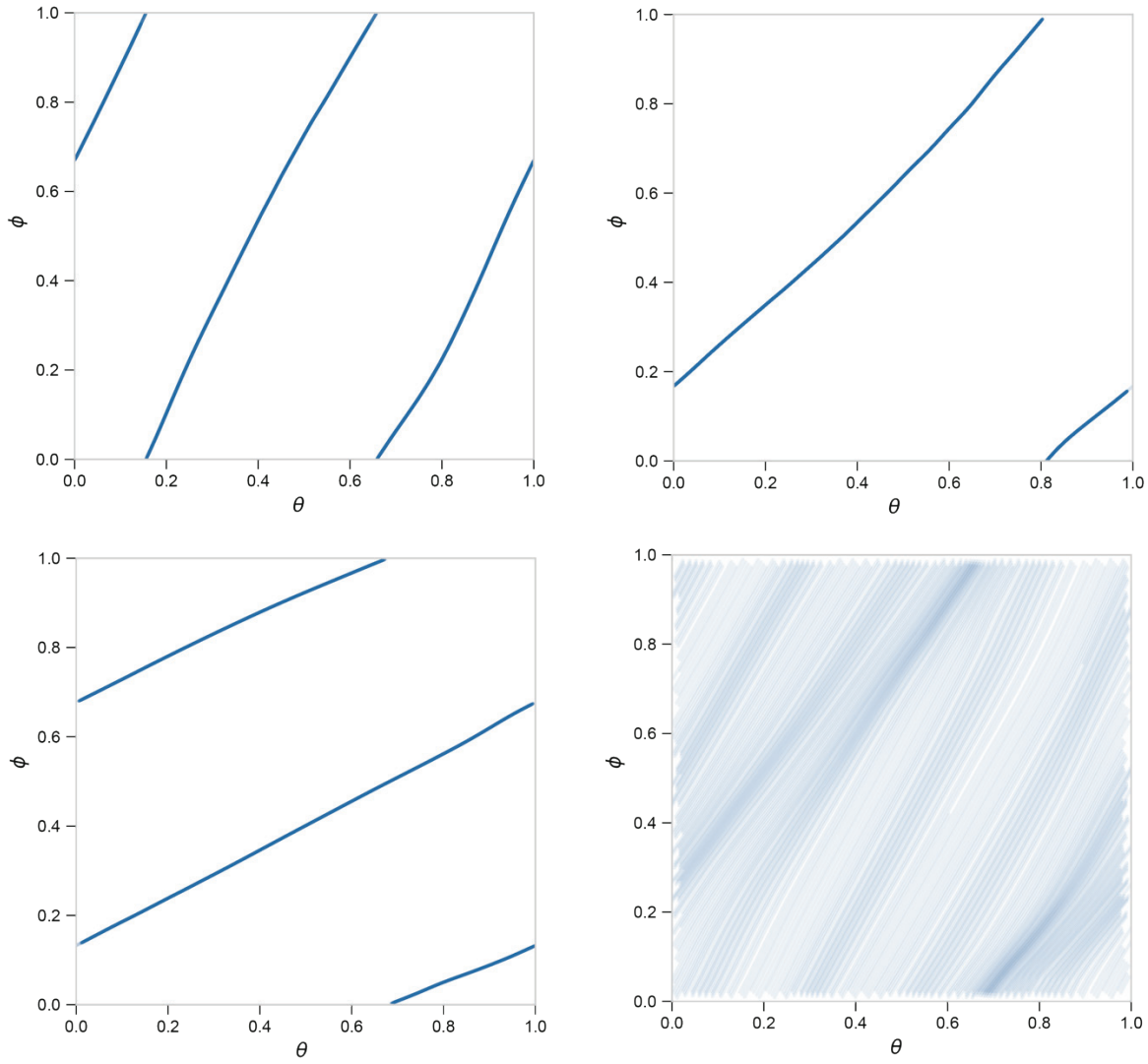
Synchronization is a fairly complex concept, and Arkady Pikovsky dedicated two whole books to try the explain the idea [1], [32]. In the simplest case, we would call *synchronized* oscillators that have agreed on a common fixed period and a fixed phase relationship. Note that the intrinsic periods of the two oscillators can be different, and so go for the phases.

However, once synchronized, the two oscillators will have the same period, such that if one looks at their phase difference at the beginning of one cycle or the next (this is called a Poincaré map, it is developed in Section 3.3.3), it should not have changed. This doesn't mean that the phase difference is constant, but instead that it also follows a periodic pattern, of the same period as those of the synchronized oscillators.

However, we could also imagine a situation where one oscillator would go exactly twice as fast as the other to which it is coupled. This is also synchronization as the phase relationship between the two oscillators also follows a periodic pattern. However, in this case, the period of the phase relationship is equal to that of the slowest oscillator.

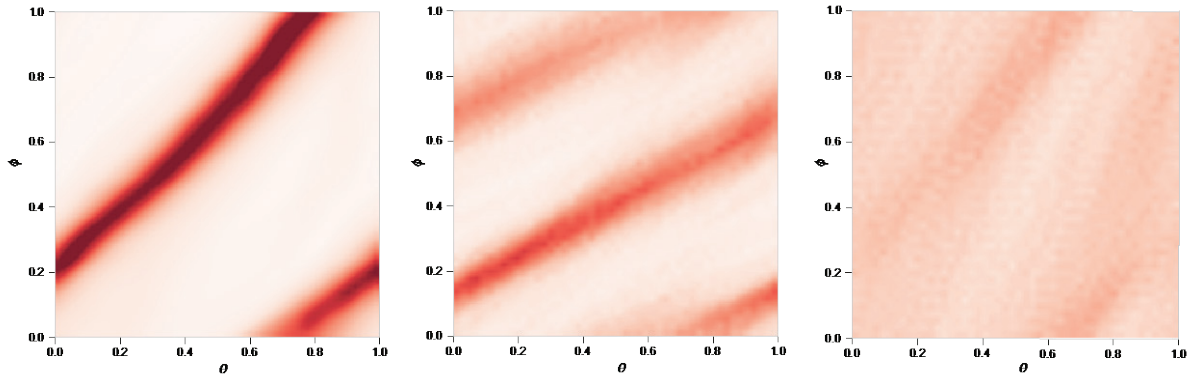
Of course, many other patterns are possible: one oscillator could go ten times as fast as the other, or, in a more complicated scenario, one oscillator could accomplish three cycles while the other do four. In practice, as long as the ratio of the two periods is a rational number, we could speak of synchronization. This phenomenon is termed $p{:}\,q$ phase-locking, where p and q denote the (integer) numbers of cycles cycle needed by the two oscillators to reach back their initial phase relationship.

Obviously, if this $p{:}\,q$ ratio were to be a non-trivial fraction (e.g. $\frac{22}{21}$), the relationship between the two oscillator phases would not be evident at all (by eye), and very long time series would be needed to recognize that there is indeed synchronization. Introduction Figure 7 illustrates different phase-lockings observed in a simulated system of two coupled oscillators. In practice, $1{:}\,1$ synchronization is far from being the rule, and complex $p{:}\,q$ phase-lockings are seen very frequently in nature. Simple examples taken from biology include the interacting cell-cycle and the circadian clock[44]–[46], or the mammalian brain[47].



Introduction Figure 7: Phase-space representation of the different types of synchronization between two oscillators having phase θ and ϕ, computed from simulations. **Top-left:** 2:1 phase-locking. **Top-right:** 1:1 phase-locking. **Bottom-left:** 1:2 phase-locking. **Bottom-right:** quasiperiodicity.

In the case of noisy systems, the $p:q$ relationship still holds, given that the noise is bounded and the coupling is large enough. However, synchronization is slightly harder to define as the two oscillators do not have precisely the same phase (each) after a full cycle, but rather the same phase plus or minus epsilon, epsilon being itself bounded[1]. If the noise is unbounded (e.g. Gaussian), true synchronization is not observed anymore as oscillators can be kicked out of equilibrium. Still, because of the coupling, some regions of the space-space are favoured over others, and therefore phase-lockings also appears on density plots; although the phases are not literally locked, as there's a continuum between synchronization and quasiperiodicity (Introduction Figure 8).



Introduction Figure 8: Phase-space representation of a system of two interacting noisy oscillators, with phase θ and ϕ computed from simulations. Noise is Gaussian. The figure shows how the system can appear to exhibit phase-locking (left, middle) or quasiperiodicity (right) by looking at the corresponding phase density. However, the concept of synchronization does not literally hold here as the noise is unbounded.

Another interesting feature of noisy systems lies in their capacity to jump between different types of phase-lockings. Indeed, several stable phase-locking modes can co-exist in the system, and the noise can kick the phase from one to the other (cf. Section 3.4.1.2). Similarly, the noise can kick the phase strongly enough that it overpasses the coupling attracting region and directly skip one cycle; this phenomenon is called phase-slipping[1].

2.5. Phase-response Curve

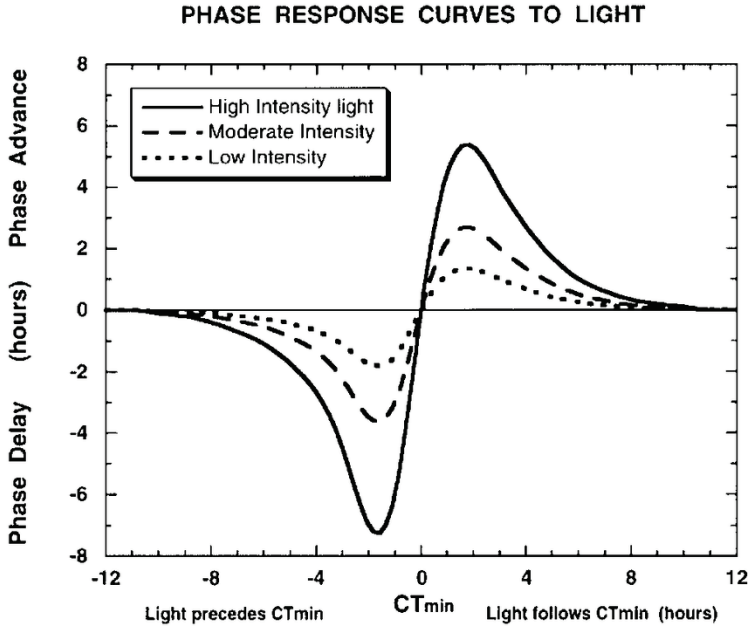
Oscillators can respond to transient perturbations in a variety of different ways. A Phase Response Curve (PRC) describes the phase response of the oscillator occurring in a cycle period depending on its current phase. It's important to mention that, in their stricter definition, PRCs are useful to describe systems response to pulse-like couplings only. They have, however, been extended to handle continuous coupling, under the name of Infinitesimal response curves (IRCs)[48]. PRCs and IRCs are interesting objects as both single and continuous pulses caninduce entrainment when applied periodically.

The principle is as follow: a single perturbation will lead to phase and amplitude change in the receiving oscillator. If the oscillator is stable, the amplitude will progressively return to its previous value, but depending on the nature of the stimulus, the phase shift may persist. One can thus track how the phase changes depending on the time and nature of the pulse. For many oscillators, the observed phase shift will greatly depend on the state of the

oscillator within the cycle (e.g. the time of the day for the circadian clock. PRCs precisely represent the phase response of the oscillator to a stimulus depending on its phase[7]).

Interesting patterns, which often have an evolutionary reason, can be observed. For instance, the human circadian clock is such that it tends to adapt its phase to the external light (Introduction Figure 9). When it's the early morning, a bit before dawn, a light pulse received by the eye will tend to advance the phase. Conversely, when it's late, light tends to slow down the phase. Without this phase-response, one could not travel far without having its external clock completely desynchronized with the actual solar hour. This is, however, not perfect, and the time delay needed for the clock to adjust is commonly known as jet-lag.

In mathematical terms, A PRC can be understood as a mapping between an old and new phase. This is not very useful in the case of single-pulsed perturbations, but continuous couplings can often be approximated by a series of single pulses. In this case, the PRC mapping may be applied several times in a row to study the phase evolution in time: this is called an *iterated map*. As explained in Section 3.3.3, synchronization can be related to the stable fixed point of the iterated map.



Introduction Figure 9: Representation of different human phase response curves, depending on the intensity of the photic cue received by the eye. CT stands for Circadian Time, where 0 is the initiation of activity in a diurnal organism (early morning for humans). Taken from *Principles and Practice of Sleep Medicine, 2017*[49].

[7] The mathematical concept of isochrons is needed to understand *why* the system reacts in such a way, from a dynamical systems perspective. This is described in Section 3.3.4.

3. Oscillators as mathematical objects

3.1. Introduction: cycling systems

Even if the term "oscillator" is now used in everyday language, it primarily refers to a physical phenomenon or an abstract mathematical object. The following section provides the mathematical framework used to model and understand the wide variety of temporal oscillators that (conceptually) exist in nature. Spatio-temporal oscillations are also widespread, but their behaviour can be complex and is outside of the scope of this thesis. Unless it is in bold, any mathematical variable will refer to a scalar. Vector will be lowercase bold, and matrices will be uppercase bold. As a rule, $\boldsymbol{t}$ will be used to refer to time.

The simplest but still general model for an oscillating system $\boldsymbol{x} \in \mathbb{R}^n$ is a differential equation of the following form:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\lambda}, t) \qquad (i)$$

$\boldsymbol{f}$ is a vectorial function that describes how the $\boldsymbol{n}$ components of the system state $\boldsymbol{x}$ change with time. This change depends on a set of parameters $\boldsymbol{\lambda}$, and, in this case, also explicitly depends on time $\boldsymbol{t}$ (e.g. the time dependence could represent a pulsed coupling, occurring only at given times). In the following section, we assume that $\boldsymbol{f}$ is of class $\boldsymbol{C^1}$, that is, it's continuous and has continuous derivatives.

Often, and as it will be the cases in the studies presented in this thesis, systems can be studied in isolation. In this case, their behaviour becomes autonomous; that is, they lose their explicit dependence on time:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\lambda}) \qquad (ii)$$

We're here interested in oscillators, that is, in systems that show sustained, stable, oscillations. Such behaviour can be described as trajectories with the following properties:

$$\begin{cases} \boldsymbol{x}(t) = \boldsymbol{x}(t+\tau) \\ \boldsymbol{x}(t') \neq \boldsymbol{x}(t) \text{ for } t < t' < t+\tau \end{cases} \qquad (iii)$$

Eq. (iii) describes how the system returns to its original state every τ units of time[8] and holds for all $\boldsymbol{x}$ components. This means that the system draws a closed trajectory $\boldsymbol{\Gamma(t)}$ in the

[8] Since the system is deterministic and autonomous, ensuring this condition for one oscillation necessary yields sustained oscillations for any time period.

n-dimensional space, every τ units of time. This trajectory can be obtained by integrating Eq. (ii)[9], and is called a limit-cycle. A stable limit cycle corresponds to a stable, strictly periodic attractor, in the sense of dynamical systems' theory[1]. This means that local perturbations will relax exponenetially fast to the periodic orbit

Depending on the nature of the system, oscillations can depend on the system's initial state or the value of the parameters. Section 3.3.2 describes how systems such as those described by Eq. (ii) can adopt limit-cycles by undergoing Hopf instabilities, which can themselves be stable (supercritical Hopf) or unstable (subcritical).

The progression of the system state along the limit-cycle is quantified by a scalar coordinate variable called the system's phase. In theory, the phase definition is arbitrary, but, for stably oscillating systems, it is usually taken as proportional to the time progression with respect to a full period. That is, denoting the phase by θ:

$$\theta(t) = 2\pi\frac{t}{\tau} \qquad (iv)$$

This means that, by construction, we set the phase to grow uniformly in time (in the direction of the motion), with frequency $\omega = \frac{2\pi}{\tau}$, considering the full cycle to be 2π long. In practice, that corresponds to a transformation from a cartesian coordinate system to a polar coordinate system which is well defined in a neighborhood of the limit-cyle:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\lambda}) \quad \Rightarrow \quad \begin{cases} \dfrac{d\theta}{dt} = \omega(\boldsymbol{\lambda}) \\ \dfrac{d\boldsymbol{r}}{dt} = \boldsymbol{g}(\boldsymbol{r}, \theta, \boldsymbol{\lambda}) \end{cases} \qquad (v)$$

Where $\boldsymbol{r}(t) \in \mathbb{R}^{n-1}$ is a coordinate perpendicular to the limit-cycle, whose behaviour is governed by the function $\boldsymbol{g}(t)$. Since the angular speed doesn't depend on time or phase, the phase is said to be a neutrally stable variable. In perturbed systems, the phase can thereby be defined outside the strict limit-cycle using the concept of isochrons (cf. Section 3.3.4), making PRCs and alike easily geometrically interpretable.

Phase equations are relevant to the study of biological oscillators, as they can hold even when the system is under small perturbations (if the oscillators are suffisciently stable and the transverse dynamics is negligible). For instance, a noisy oscillator having phase θ entrained by a periodic forcing with phase φ can be described with an equation of the form (cf. the project presented in Chapter 1):

[9] This is often hard to do analytically, but numerical integration will be tractable for most systems.

$$\frac{d\theta}{dt} = h(\theta, \varphi) + \varepsilon \qquad (vi)$$

Similarly, a noisy oscillator whose noise is phase dependent will follow a stochastic ODE of the form:

$$\frac{d\theta}{dt} = \omega + \varepsilon(\theta) \qquad (vii)$$

3.2. Conditions needed for oscillations

3.2.1. Introduction

Although the previous section explains how oscillations can be described by dynamical systems, it doesn't explain what physical or chemical mechanisms lead to their emergence in the first place. Said differently, what kind of system structural design leads to oscillations? In the literature[25], [50], the following requirements are usually considered as the necessary (but not sufficient) basis for generating oscillations:

- first, an inhibitory feedback loop, which includes one or more oscillating variables, is needed to carry the system back to the starting point of its oscillation.
- Then, there must be a source of delay in this feedback loop, which allows an oscillating variable to overshoot a steady-state value before the feedback inhibition is fully active.
- Finally, the governing equations for the components at play must be sufficiently nonlinear to destabilize the steady-state, but chosen such that the producing/degrading equations occur on appropriate time scales that permit the system to generate oscillations.

Without negative feedback loop(s) and delay, there can't be oscillations, as the system components would either diverge in time, or stabilize if there is some kind of active/passive degradation in the system. However, negative feedback alone is not sufficient to generate oscillations, as, without further perturbation, the system would instead monotonically return to steady state[4]. It is therefore needed to add some delay in the system, such that the oscillating variables overshoot the steady-state before the feedback truly comes into play. Delay can be explicitly added to the system using delay differential equations[51], but an alternative way to model it consists in adding components in the negative feedback loop in order for it to take more time to be effective. In practice, in biochemical systems, delay often stems from spatial aspects of the system: molecules must diffuse in cytoplasm and through membranes[52]. In population dynamics, space is also involved, but delay can also come from

the fact that individuals can take time to die of starvation when no prey/nutrients are available, or conversely to reproduce when there's more food than needed[53]. Finally, non-linearity is usually considered essential to generate sustained oscillations. In practice, oscillations can be obtained from linear systems (harmonic oscillators, which behave as neutrally stable *centres*, cf. Section 3.2.2.2), but limit-cycles as introduced in Section 3.1 are inherently nonlinear phenomena, with oscillations determined by the structure of the system itself. Non-linearity enables to change the behaviour of the system in a way that depends on its own state. A linear system can generate exponentially growing or decaying oscillations (spirals), but adding a non-linear repression term can stabilize these oscillations to a stable value.

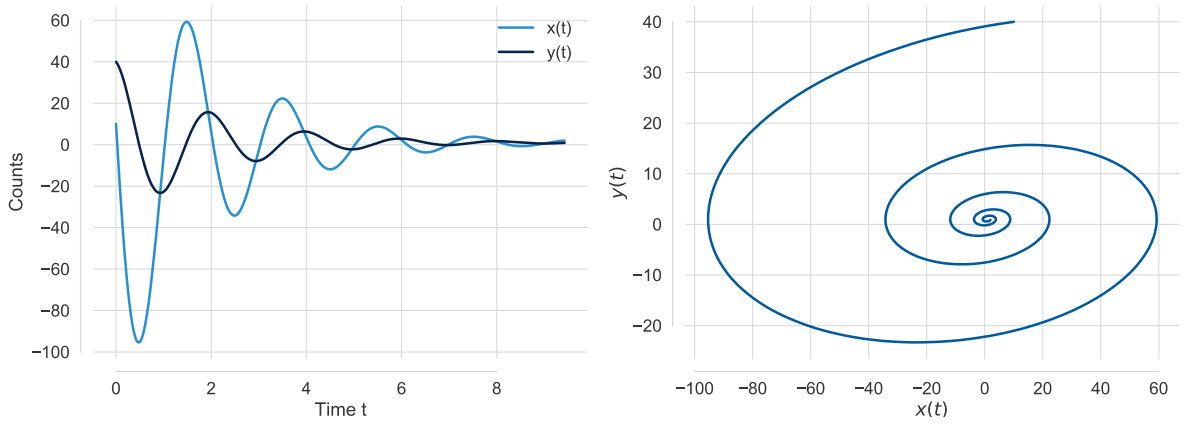
3.2.2. A few introductory examples

3.2.2.1. A damped linear oscillator

The simplest way to model oscillations with an ODE system is to consider one specie x who is produced at constant rate α and gets degraded through another specie y, which is itself produced at a rate x, and degraded proportionally to its own value (that is, we have a negative feedback loop made of two arrows, as x is used to produce y, which represses x production):

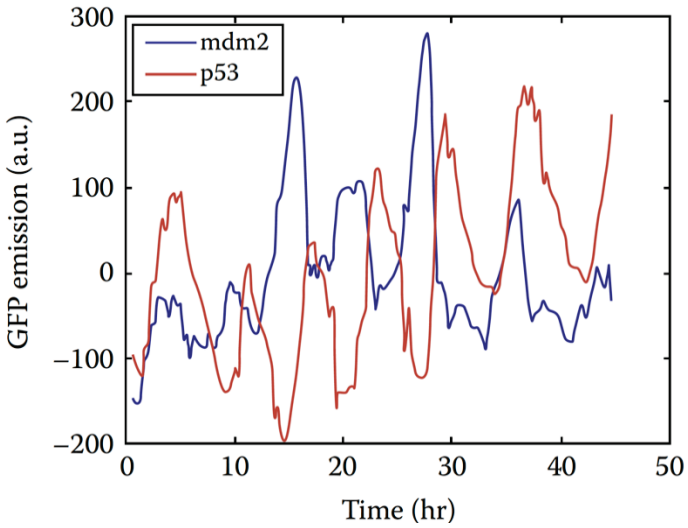
$$\begin{cases} \dot{x} = \alpha - \beta y \\ \dot{y} = x - y \end{cases} \qquad (viii)$$

Simulating the system with an appropriate timescale for the different parameters yields damped oscillations (Introduction Figure 10).



Introduction Figure 10: Simulation of a damped oscillator. **Right:** temporal trajectories of the two oscillators components x and y. **Left:** phase-space representation of the temporal trajectories on the left.

Now, 2D noiseless linear system can't generate asymptotically stable sustained oscillations[10][2], but biological systems are often very noisy. In practice, noise kicks the system away from its spiralling stable fixed point and prevents the oscillations from damping out. This kind of noisy oscillations was observed in the interacting system of p53 (a protein that prevents DNA damage) with mdm2 (which plays the role of negative feedback for p53)[54], as illustrated Introduction Figure 11 below:



Introduction Figure 11: Temporal oscillations of the proteins p53 and mdm2. This system can be shown to be a kicked damped oscillator, as visible by the variable amplitude between cycles. Figure taken from the book *An introduction to Systems Biology* by *Uri Alon*[4].

Noise-induced oscillations are readily identifiable as the amplitude A of the pulses follow a law of the type: $P(A) \sim A e^{\frac{-A^2}{A_0^2}}$ [55], [56]. However, in many biological systems, Evolution has not favoured this type of oscillations as amplitude and frequency must be tightly controlled, and noise is often very dependent on the external conditions (e.g. temperature, pressure, etc.).

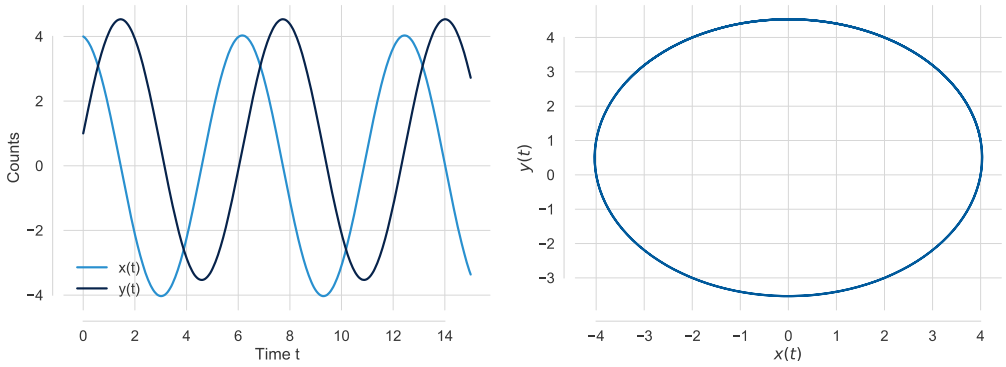
3.2.2.2. A harmonic oscillator

The simplest example of a system generating sustained oscillations is the harmonic oscillator. Although it originates from a real physical system (a mass hanging from a spring), it can be written generically as the following linear ordinary differential equation (ODE) model:

$$\begin{cases} \dot{x} = y \\ \dot{y} = -\omega^2 x \end{cases} \qquad (ix)$$

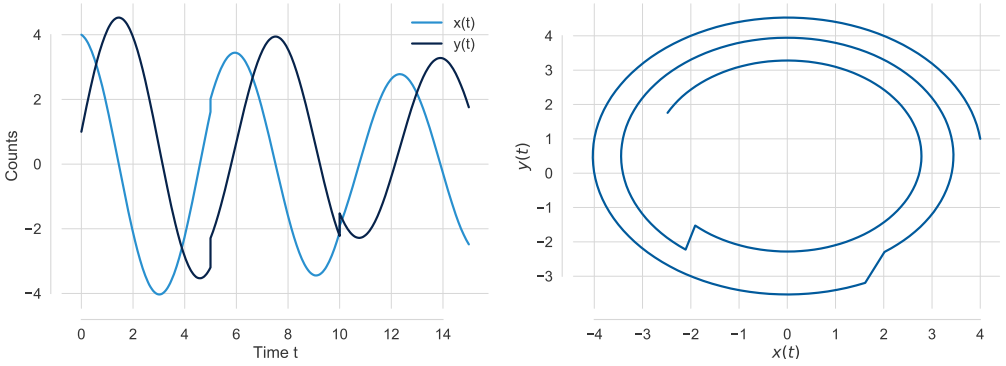
[10] This is because stable limit-cycles require non-linearities.

Simulating this system yields regular oscillations, as illustrated in Introduction Figure 12 below:



Introduction Figure 12: Simulation of a harmonic oscillator. **Left:** Temporal oscillations observed for the two components of the oscillator. **Right:** Representation of the oscillator trajectory in the phase-space of the system.

The solution of Eq. (ix) can easily be found to be $x(t) = A\cos(\omega t + \varphi)$, where A, φ will depend on the initial condition. In practice, this type of oscillator is rarely encountered in biological systems. First, because biochemical reactions are themselves non-linear, and then because Eq. (ix) corresponds to a conservative system, in which both amplitude and phase depend on the noise and the initial condition (Introduction Figure 13, random kicks have been added to the simulation).



Introduction Figure 13: Simulation of the same system as in Figure 11, using the same representations, with the exception that the system is now perturbed at time $t = 5$ and time $t = 10$, yielding persistent changes in phase and amplitude due to the energy conservative nature of the system.

Finally, such an oscillator wouldn't be temperature compensated. Therefore, it could not act as a clock.

3.2.2.3. Negative feedback oscillators

System with feedback loops can harbour sustained, undamped oscillations (stable limit cycles in the above sense, cf. Section 3.1) which are also robust to noise. That is, the more components there are in the cycle, the more delay is added, and the easier it is to find a range for parameters yielding oscillations. Such a system is called a delay oscillator.

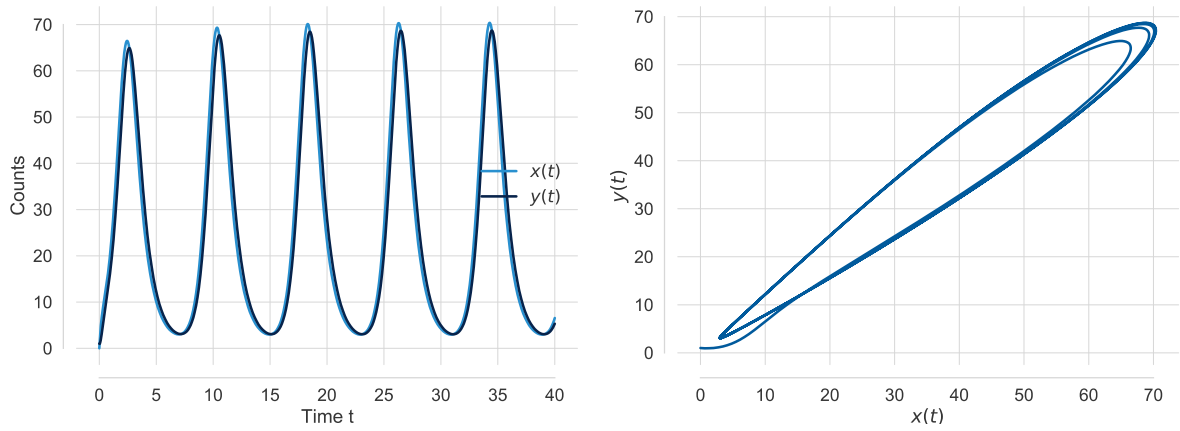
In practice, negative feedback oscillators can also be modelled by lower dimensional systems of equations, in which the negative feedback act through an explicitely introduced delay, and the system takes then the form of a Delay Differential Equation (DDE). I will not further diuscuss DDE's in this thesis.

One of the most famous negative feedback oscillators probably originates from the work of Brian Goodwin[59], in which he theorized a cycle of three repressors, that is now known as the Goodwin-Griffith three-variable model. Alternative version of this model have notably been used to model the circadian clock[60]. In its simplest form, the corresponding system of equations is as follow:

$$\begin{cases} \dot{x} = -\alpha_1 + \dfrac{K^n}{K^n + z^n} - \gamma_1 x \\ \dot{y} = \alpha_2 x - \gamma_2 y \\ \dot{z} = \alpha_3 y - \gamma_3 z \end{cases} \qquad (x)$$

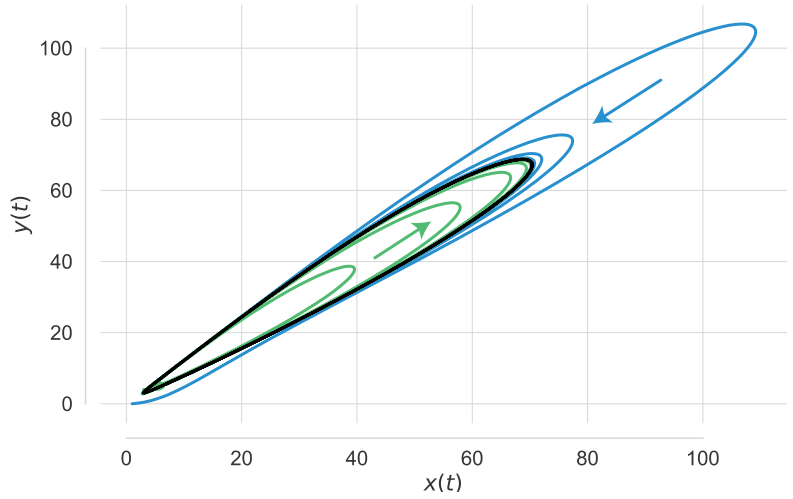
This model can be understood as follow: a given gene codes for a mRNA x, which is translated into a protein y. This protein activates an inhibitor of x, called z. This repression is described by a nonlinear, hyperbolic function, which decreases with increasing inhibitor concentration and determines the transcription rate. The first-order degradation rates γ_i ensure that the variables remain positive, and the delay involved through the use of z to repress x favours the occurrence of self-sustained oscillations.

Simulating the system with the proper set of parameters yields regular temporal oscillations, corresponding to a closed trajectory in the phase-space: a limit-cycle (Introduction Figure 14).



Introduction Figure 14: Temporal (**left**) and phase-space (**right**) representation of the components x and y taken from a simulation of the Goodwin-Griffith three-variable model. A limit-cycle is observed, as demonstrated by the presence of sustained oscillations in the temporal trajectories, and closed orbit in the phase-space.

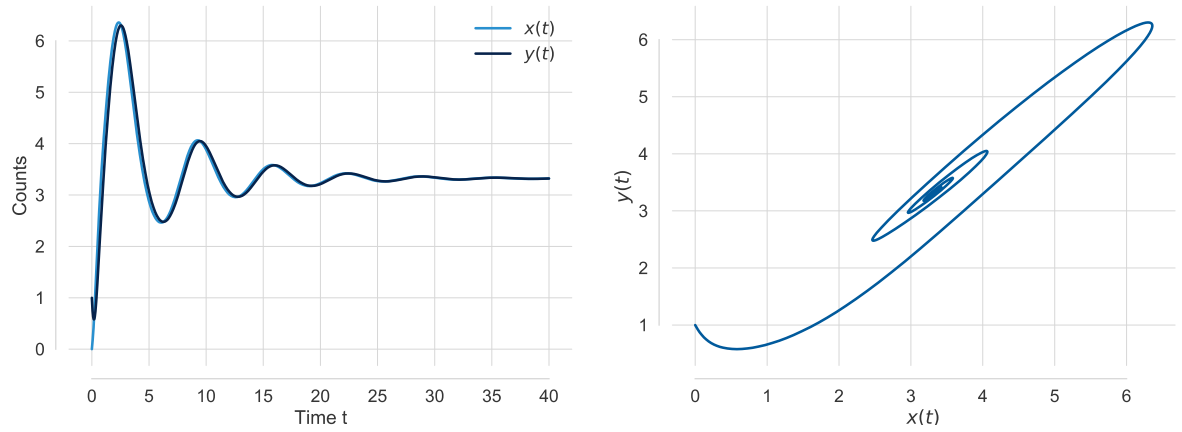
Now, the interesting property of this system is that, contrary to the harmonic oscillator presented above, the oscillations are relatively[11] resilient to noise and initial condition, as shown Introduction Figure 15 below:



Introduction Figure 15: Phase-space representation of the Goodwin-Griffith three-variable model simulated as in Figure 13. Two different initial conditions have been taken (blue and green), both leading to the same limit-cycle (black) after a few oscillations.

Another interesting observation is that, depending on the value of the parameters, the limit-cycle can disappear and collapse into a stable spiral (Introduction Figure 16).

[11] In most cases, the stable limit-cycle will be surrounded by a given attracting pool. Outside of this pool, trajectories will diverge or tend towards other fixed points. Therefore, the noise must not push trajectories outside of the attracting pool.



Introduction Figure 16: Temporal (**left**) and phase-space (**right**) representation of the components x and y simulated as in Figure 13, but with a different set of parameters. No stable limit-cycle exists in the system anymore, as shown by the presence of damped oscillations in the temporal trajectories, and a stable spiral in the phase-space.

This phenomenon is well known from dynamical systems as a Hopf bifurcation. It explains how, depending on the eigenvalues, the system can exhibit sustained or damped oscillations. This is developed in Section 3.3.2.

3.2.3. Conclusion

In practice, many structural designs can lead to sustained oscillations. For instance, adding positive feedback in addition to the negative one can also add delay to the system, increasing the robustness of the oscillatory behaviour (this is what is observed in relaxation oscillators[4]). Autoregulation can also be used in various ways to tune the frequency of the oscillations. Finally, nonlinearity itself can come in many forms: a Hill function in the Goodwin Model, but also in the famous repressilator developped by Elowitz and Leibler[61], quadratic functions in the Brusselator (used in biochemistry to model autocatalytic reactions)[62], cubic function in the Van del Pol oscillator[63], etc.

Unfortunately, biological systems comprising more than three or four interacting species are often intractable analytically because of the multitudes of factors that come into play to generate the observed behaviour. From a bottom-up approach, numerical simulations can be used to see which type of behaviour emerges from which structure. From a top-down approach, if these systems exhibit sustained oscillations, their dynamics can be approximated by a phase oscillator representing the progression along the corresponding limit-cycle.

3.3. Oscillators in dynamical systems theory

I hereafter introduce cycling systems (cf. Section 3.1) from a dynamical system's point of view. Dynamical system's theory is a relatively recent branch[12] of mathematics and physics that has its origins in Newtonian mechanics, but happens to describe very well the behaviour of oscillating systems. Eq. (xi) is the prototypical forms of a dynamical system, where the time-evolution of the set of points $\boldsymbol{x}$ is governed by an ODE (here autonomous). In general, we'll be interested in systems of the form:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n \qquad (xi)$$

Note that, for simplicity, we don't represent the set of parameters $\boldsymbol{\lambda}$ in this equation. They are implicitely contained in $\boldsymbol{f}$. In addition, we only consider autonomous dynamical systems.

3.3.1. Linear stability analysis

Without simulations, the behaviour of a non-linear system is usually hard to predict. However, it can be well approximated around the fixed points of the system, that is, the points $\boldsymbol{x}^*$ such that $\frac{d\boldsymbol{x}^*}{dt} = 0$. This approximation is obtained by linearizing the system in the neighbourhood of the fixed point, and then applying a method called stability analysis. Linearizing the function from Eq. (xi) yields:

$$\boldsymbol{f}(\boldsymbol{x}) \underset{\boldsymbol{x}=\boldsymbol{x}^*}{\approx} \boldsymbol{f}(\boldsymbol{x}^*) + \nabla \boldsymbol{f}|_{\boldsymbol{x}^*} \cdot (\boldsymbol{x} - \boldsymbol{x}^*) \qquad (xii)$$

In this equation, $\nabla \boldsymbol{f}|_{\boldsymbol{x}^*}$ is the Jacobian of $\boldsymbol{f}$ evaluated at the fixed point $\boldsymbol{x}^*$. Calling f_i and $x_i, i \leq n$, the components of $\boldsymbol{f}$ and $\boldsymbol{x}$, we get:

$$\nabla \boldsymbol{f}|_{\boldsymbol{x}^*} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \qquad (xiii)$$

Now, in most cases[13], this matrix can be diagonalized:

$$\nabla \boldsymbol{f}|_{\boldsymbol{x}^*} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} \qquad (xiv)$$

[12] The birth of dynamical systems theory is usually attributed to Henry Poincaré, who laid its foundations at the end of the XIXth century.

[13] This can happen if the algebraic and geometric multiplicities of at least one eigenvalue do not coincide (e.g. a 2-dimensional system with a 1-dimensional eigenspace). These cases can still be solved[2].

Where $\boldsymbol{P}$ is the transfer matrix, having the eigenvectors of $\nabla \boldsymbol{f}|_{\boldsymbol{x}^*}$ as columns:

$$\boldsymbol{P} = (\boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_n) \qquad (xv)$$

And $\boldsymbol{A}$ is the diagonal matrix containing the eigenvalues of $\nabla \boldsymbol{f}|_{\boldsymbol{x}^*}$:

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \qquad (xvi)$$

Now, by construction, for each eigenvalue, we have:

$$\mathbf{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i \qquad (xvii)$$

Since the eigenvectors are linearly independent, it can be shown that the general solution for $\boldsymbol{x}$ is simply[14]:

$$\boldsymbol{x}(t) = \sum_{i=1}^{n} c_i e^{\lambda_i t} \boldsymbol{v}_i \qquad (xviii)$$

It should be quite clear from this equation than if at least one eigenvalue is positive, the system is going to (locally) diverge. Conversely, if all eigenvalues are negative, the system is going to tend towards the fixed point $\boldsymbol{x}^*$. In practice, we're interested in oscillations, and oscillations can be shown to exist when eigenvalues are complex[15]. Indeed, let $\lambda = a + i\omega$. According to Euler's formula:

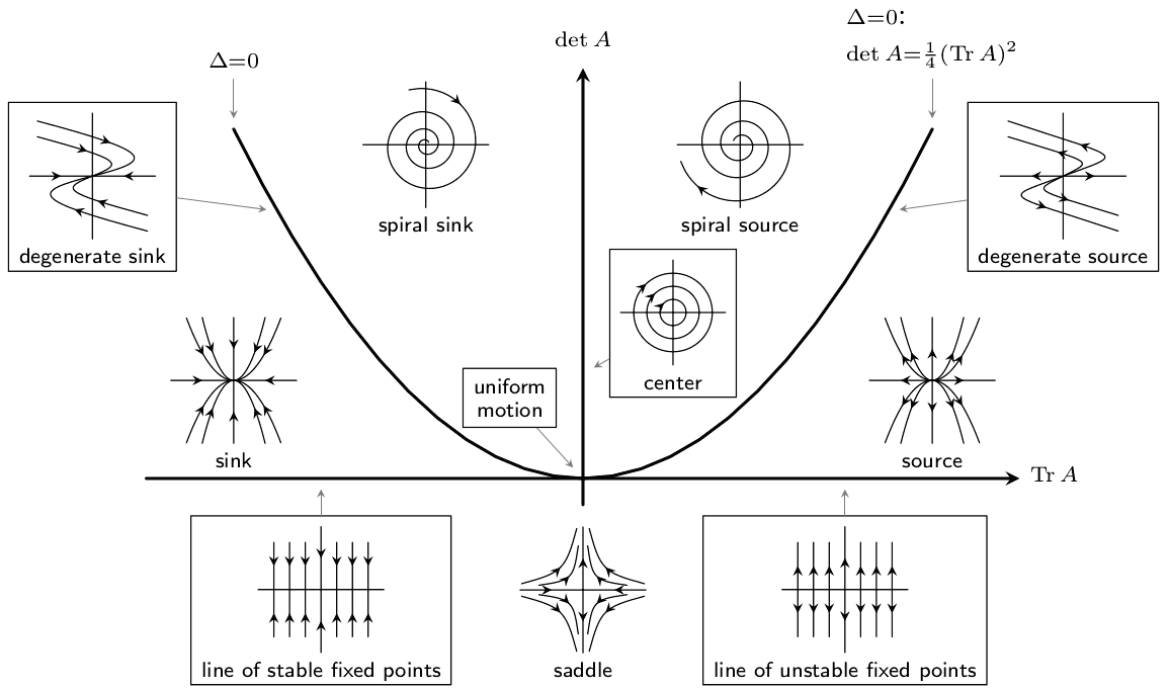
$$e^{\lambda t} = e^{(a+i\omega)t} = e^{at}[\cos(\omega t) + i \sin(\omega t)] \qquad (xix)$$

Thus, the general solution will be a combination of harmonic functions and will be periodic itself. Now, in a linear system, oscillations can either grow in time ($\mathcal{R}e(\lambda) > 0$, this is called unstable spiral), either decay in time ($\mathcal{R}e(\lambda) < 0$, stable spiral), either stay constant ($\mathcal{R}e(\lambda) = 0$, centres). However, none of these cases yields stable oscillations in noisy systems. In 2-dimensional systems, the whole variety of behaviours observed, including oscillating ones, can be summarized according to the trace and determinant of the system[16], as shown in Introduction Figure 17.

[14] This can be verified easily by plugging the solution in Eq. (xi)

[15] Note that complex eigenvalues are always conjugate, so are eigenvectors.

[16] The relative value of trace and determinant is used instead of the eigenvalues themselves as it is faster to compute (no diagonalization is required) and, in two dimensions, it can exhaustively summarize the possible topologies.



Introduction Figure 17: Stability diagram classifying two-dimensional dynamical systems as stable or unstable according to their trace and determinant. Taken from Wikimedia Commons.

3.3.2. Hopf bifurcation and Limit-cycle

Self-sustained, stable oscillations can always be shown to originate from an attracting limit-cycle. The general parametric conditions needed for a limit-cycle to emerge are well defined: this the theory of Hopf bifurcation[17]. There are several types of Hopf bifurcations (subcritical, supercritical, degenerate), but all of them explain the same phenomenon, that is, how a limit-cycle (stable or unstable) appears as the eigenvalues of a system increases and cross the imaginary axis, depending on the value of a control parameter.

Let's consider an elementary example, analyzing the following system in polar coordinates:

$$\begin{cases} \dot{r} = r(\alpha - r^2) \\ \dot{\theta} = \omega \end{cases} \qquad (xx)$$

Clearly, if $\alpha < 0$, then $\dot{r} < 0$ and so this system is going to tend to a stable fixed point (spiralling, since $\dot{\theta} \neq 0$). Now, if we progressively increase the value of α, it will become

[17] Limit-cycle can actually appear from another type of bifurcation called global bifurcation. This type of bifurcation is harder to detect as it involves large regions of the phase-space, but the overall concept is the same: a limit-cycle appear as a control parameter is varied[2]. Limit-cycle can also appear if an artificial resetting is forced onto the system, as in the case of SNIC oscillators[64].

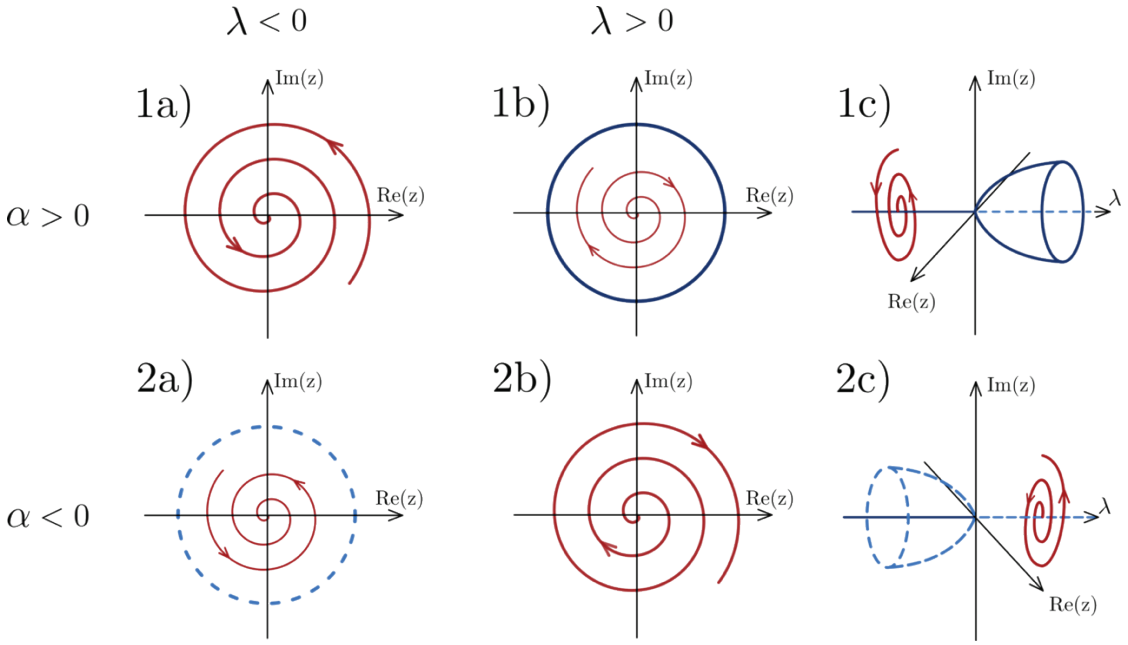
positive at some point, and it can easily be shown that the system is going to stabilize at $r = \sqrt{\alpha}$: a limit-cycle just appeared.

Now, this system is a trivial case as it is parametrized in polar coordinates, and the oscillating behaviour is uncoupled from the radius. However, the general principle applies to any system: a limit-cycle will always appear as a control parameter passes a given threshold.

In practice, it can be shown that any system going through a Hopf bifurcation can be approximated by the following equation around equilibrium:

$$\dot{z} = z\big((\lambda + i) + (\alpha + i\beta)|z^2|\big) \qquad (xxi)$$

This is the canonical form of a Hopf bifurcation[65], in which z is complex numbers, while λ, α, β are real. This equation summarizes the two interesting behaviours of Hopf bifurcations, illustrated Introduction Figure 18 below. If $\alpha < 0$, then it can be shown that there exists a stable limit-cycle when $\lambda > 0$: this is a supercritical bifurcation. Conversely, if $\alpha > 0$, there is an unstable limit-cycle when $\lambda < 0$: this is a subcritical bifurcation.



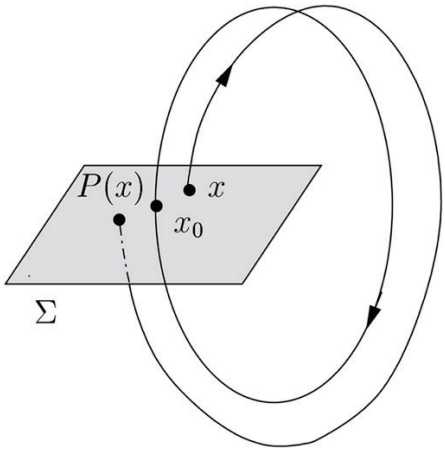
Introduction Figure 18: Dynamics of the Hopf bifurcation near $\lambda = 0$. Possible system trajectories in red, stable structures in dark blue and unstable structures in dashed light blue. (**top**) Supercritical Hopf bifurcation, in which the phase space dynamics goes from a stable spiral (**a**) to unstable spiral surrounded by a stable limit-cycle (**b**), depending on the value of λ (**c**). (**bottom**) Subcritical Hopf bifurcation, in which the phase space dynamics goes from a stable spiral surrounded by an unstable limit-cycle (**a**) to an unstable spiral (**b**), depending on the value of λ (**c**). Figure and caption adapted from Wikipedia[66].

3.3.3. Poincaré map

The Poincaré map is a handy tool to study oscillators properties and phase synchronization. The principle is to study the evolution of a system in a given location in space, as if only a tiny fraction of it was visible.

Consider a system with coordinates x showing a limit-cycle in dimension n. Now let Σ be a transverse section of this system, in the form a hyperplane in $n-1$ dimensions. A Poincaré map is defined as a function $P(x)$ that maps Σ to itself, obtained by following $x(t)$ from one intersection with Σ to the next. That is, if x_i denotes the i^{th} intersection, the mapping is defined as $x_{i+1} = P(x_i)$ (Introduction Figure 19).

Now, in the presence of a stable limit-cycle, trajectories should be perfectly closed, so $P(x_{i+1}) = P(x_i), \forall i$. This means that there must exist fixed point x^* on Σ such that $P(x^*) = x^*$, belonging to every trajectory. This proves the existence of a closed orbit for the dynamical system under study.



Introduction Figure 19: Cartoon representation of the Poincaré mapping principle, in which a hypersection Σ (grey plane) is crossed by a cycling trajectory (black line). The Poincaré mapping is such that $x = P(x_0)$. Taken from *Brachtendorf et al., 2014*[67].

Now, the exciting part is that, by studying the behaviour of $P(x)$ near the fixed point, one can find out the stability of the closed orbit. That is, depending on the sequence of elements generated by P (convergent or not), one can decipher if there's indeed a stable limit-cycle in the system. The difficulty of such an approach is that it is usually tricky, or even impossible, to find an analytical form of P. However, the poincaré mapping can still be studied numerically, e.g. to reveal how fast a system is diverging from an unstable fixed point in time.

There exists a particular version of the Poincaré map called the Stroboscopic map, in which the behaviour of the system is recorded every T units of time. That is, instead of looking at the system at given points in space, one takes an interest in given points in time. As before,

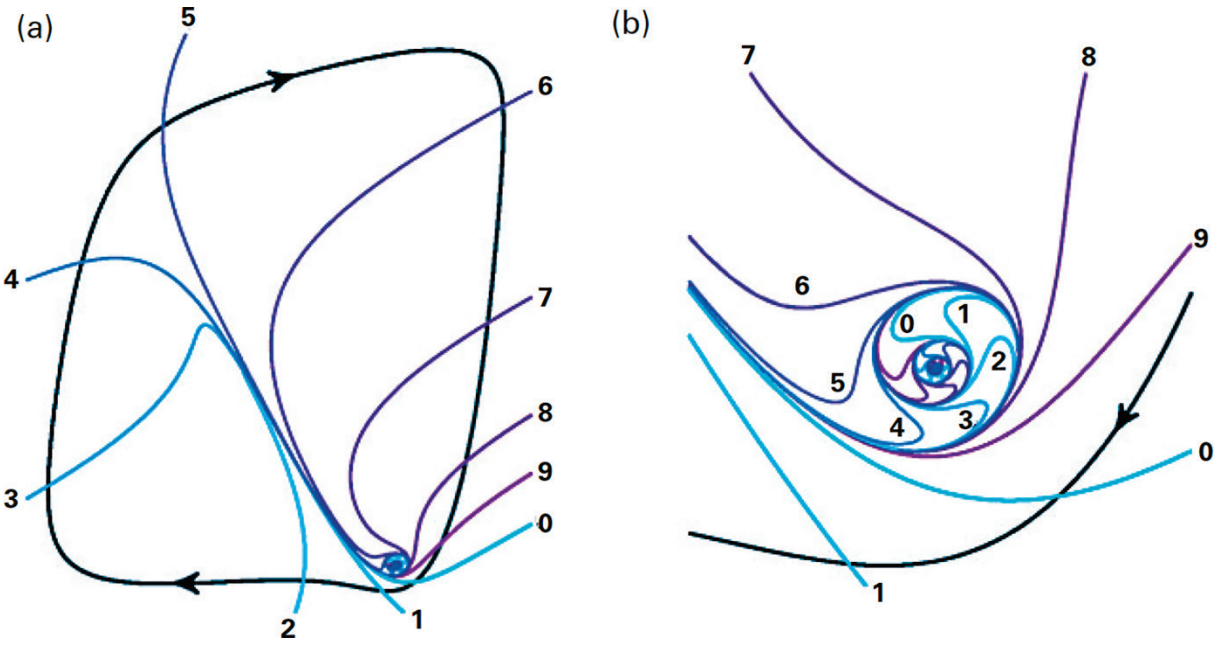
this is very useful to establish the presence of a stable limit-cycle: if the oscillator phase does not change after every mapping, that means that there exist sustained oscillations, with period T.

3.3.4. Isochrons

Determining the instantenous phase of a biological oscillator is a problem that frequently arises in practice. Indeed, such systems are usually not evolving strictly on their limit-cycle, but rather around it, being regularly pushed away from the attractor by the noise (cf. Section 3.1). *Isochrons* (Introduction Figure 20) define the phase for all the points that eventually approach the limit-cycle. The principle is the following: in a noiseless system, consider two trajectories $x_1(t)$ and $x_2(t)$ having initial conditions a_1 and a_2 in the vicinity of the limit-cycle. Then a_1 and a_2 are considered to have the same phase if the two trajectories $x_1(t)$ and $x_2(t)$ asymptotically approach each other, that is:

$$\forall \varepsilon, \exists t, t' \text{ such that } \forall t > t', |x_1(t) - x_2(t)| < \varepsilon \qquad (xxii)$$

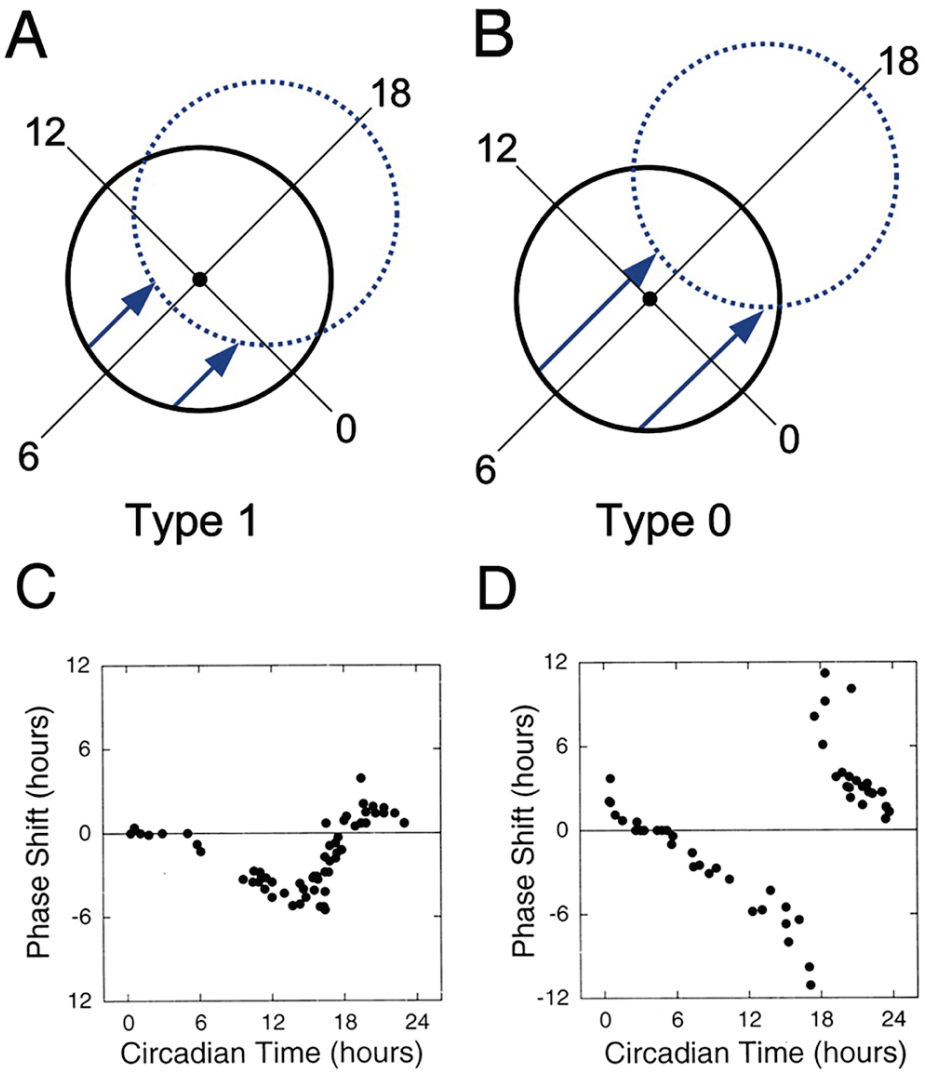
Defining phase this way allows accounting for transients, including noise-induced transients.



Introduction Figure 20: Illustration of the isochrons (coloured lines) from the Fitzhugh-Nagumo model[68] in a wide (**a**) and closed-up view (**b**) around the fixed-point of the system. While some systems show remarkably simple isochrons, this is an example of a simple system that has remarkably complex isochrons. Figure and captions adapted from the book by *Daniel Forger*[3].

Isochrons are also very useful to better characterize PRCs, which, essentially, describe how the phase of a given oscillator reacts to a given perturbation (cf. Section 2.5). For instance, PRCs can be classified into type 0 and type 1 depending on whether the corresponding perturbations are strong or weak, respectively[69] (Introduction Figure 21). There's a

topological difference between the two as strong perturbations tend to push the phase beyond the intersection of the isochrons, while weak perturbations leave the system in the close vicinity of the limit-cycle. Now, the phase is considered ill-defined at the intersection of the isochrons (usually, near the centre of the limit-cycle), such that a singularity is observed under the form a phase discontinuity in Type 0, but not in Type 1, PRCs.



Introduction Figure 21: Representation of the two types of PRC, depending on the strength of the perturbation relative to the isochrons intersection, using a limit-cycle model of the circadian pacemaker. (A-B) A limit-cycle oscillator is represented as a black (dashed blue) circle before (after) perturbation (light pulse, blue arrows). Time moves clockwise around the circle, and four phase points are indicated by the radial lines that represent the isochrons. Singularity, an unstable equilibrium point, arises at the intersection of the isochrons. In (A), the strength of the light pulse is weak, and cannot push the system across the singularity to the opposite phases of the cycle. Thus, the new phase is similar to the old phase and the resetting is Type 1. In (B), the strength of the light input is stronger and the system is pushed across the singularity, which results in very large phase shifts or Type 0 resetting. (C-D) Phase-shifting responses to 6-h light pulses in the circadian activity rhythms of C57BL/6J wild-type and Clock/+ mice, corresponding to Type 1 (C) and Type 0 (D) PRC. The x axis indicates the CT at the beginning of the light pulse. The y axis indicates the phase shift produced by the light pulse. Figure and caption adapted from *Vitaterna et. Al, 2006*[70].

3.4. Coupling and synchronization

3.4.1. Synchronization of two oscillators

3.4.1.1. Deterministic case

Consider two oscillators having phases ϕ_1 and ϕ_2, and natural frequencies ω_1 and ω_2. We assume that these oscillators are mutually influencing each other and model their interactions through two coupling terms F_1 and F_2, each depending on the value of the phase of both oscillators. We also add a term ϵ representing the coupling strength. The final system of equation is:

$$\begin{cases} \frac{d\phi_1}{dt} = \omega_1 + \epsilon F_1(\phi_1, \phi_2) \\ \frac{d\phi_2}{dt} = \omega_2 + \epsilon F_2(\phi_1, \phi_2) \end{cases} \qquad (xxiii)$$

Note that F_1 and F_2 are 2π periodic in both their arguments. We can define the phase difference as:

$$\Delta\phi = \phi_1 - \phi_2 \qquad (xxiv)$$

A condition for synchronization is $\frac{d\Delta\phi}{dt} = 0$. Less stringent conditions are also possible, leading to a periodic solution for $\frac{d\Delta\phi}{dt}$, but for now, we only introduce the simplest case.

In the case of $1:1$ synchronization, it is clear that the two natural frequencies ω_1 and ω_2 should be close to resonance, i.e. $\omega_1 \simeq \omega_2$. Expanding the functions F_1 and F_2 into Fourier Series, averaging all the fast oscillating terms (which have no impact on the synchronization) yield the one-dimensional resonant form of the coupling functions: q_1 and q_2. After some algebra, the temporal derivative of $\Delta\phi$ can be written:

$$\frac{d\Delta\phi}{dt} = (\omega_1 - \omega_2) + \epsilon\big(q_1(\Delta\phi) - q_2(\Delta\phi)\big) \qquad (xxv)$$

Defining $q(\Delta\phi)$ as $q_1(\Delta\phi) - q_2(\Delta\phi)$, it can be shown that the condition for synchronization between the two oscillators is:

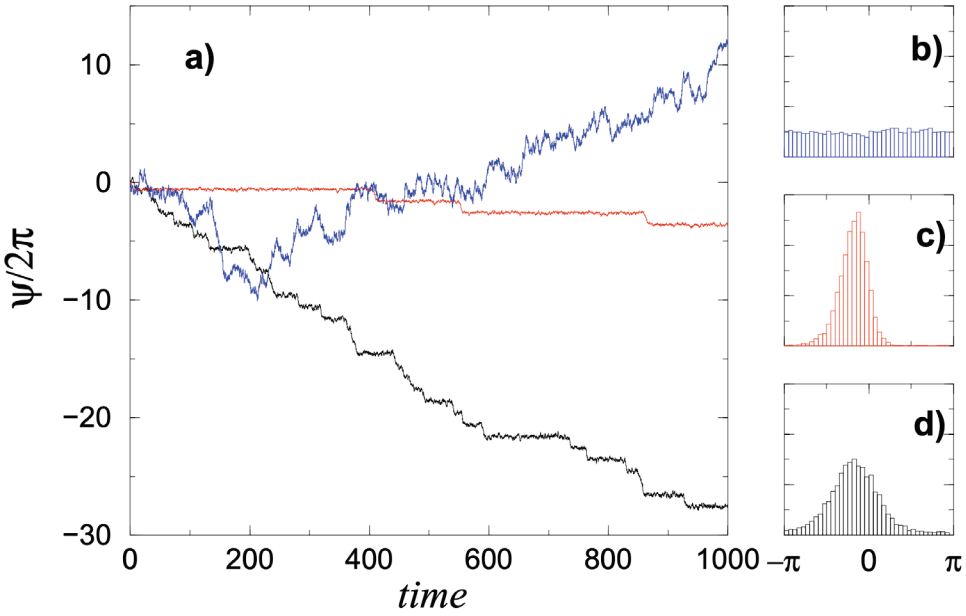
$$\epsilon q_{min} < \omega_1 - \omega_2 < \epsilon q_{max} \qquad (xxvi)$$

In this equation, q_{min} and q_{max} are resp. the minimum and maximum values of q. It follows that synchronization can occur as long as the natural frequencies of the two oscillators are not too different, and that the coupling is strong enough.

3.4.1.2. Noisy case

In the case of noisy systems, a noise term ε (e.g. Gaussian noise, here independent of the phase) is added to Eq. $(xxiii)$:

$$\begin{cases} \frac{d\phi_1}{dt} = \omega_1 + \epsilon F_1(\phi_1, \phi_2) + \varepsilon \\ \frac{d\phi_2}{dt} = \omega_2 + \epsilon F_2(\phi_1, \phi_2) + \varepsilon \end{cases} \qquad (xxvii)$$



Introduction Figure 22: (**a**) Fluctuation of the phase difference in a noisy oscillator. Without forcing, the behaviour of the ϕ is diffusive: It performs a motion that reminds a random walk (blue curve); the distribution of the $\phi \mod 2\pi$ is shown in (**b**), it is practically uniform. External forcing with nonzero detuning suppresses the diffusion, the phase of the oscillator is nearly locked (red curve), but sometimes phase slips occur; the respective distribution (**c**) becomes rather narrow and unimodal. Stronger noise (black curve) causes more phase slips, so that there are only rather short epochs where ϕ oscillates around a constant level; the distribution of the $\phi \mod 2\pi$ remains nevertheless unimodal (**d**). Figure and caption taken from *Phase Synchronization in Regular and Chaotic Systems*, *Pikovsky et al., 2000*[71].

Taking the same steps as before lead to a Langevin equation for the phase difference:

$$\frac{d\Delta\phi}{dt} = (\omega_1 - \omega_2) + \epsilon Q(\Delta\phi) + \xi \qquad (xxviii)$$

Where ξ is again a noise term, and Q is a periodic function built as a combination of the Fourier series of the coupling functions F_1 and F_2. Now, a one-dimensional Langevin dynamics can be described as a random-walk of a particle in a potential, whose equation is:

$$V(\Delta\phi) = (\omega_1 - \omega_2)\Delta\phi - \epsilon \int^{\Delta\phi} Q(x)dx \qquad (xxix)$$

Here, this means that if the coupling Q is strong enough to compensate the frequency difference $(\omega_1 - \omega_2)$, the potential V will have substantial barriers, and the phase-difference will tend to remain constant. If not, the noise will kick the particle out of the potential, and phase-slips will be observed. In any case, the standard deviation of the density curve around the attractor is proportional to the noise. This phenomenon has been beautifully explained in reference [71], from which [70] below is taken.

3.4.1.3. Other types of synchronization

As explained in Section 2.4, different types of synchronization exist. For instance, one could perfectly imagine a system in which the time needed for one oscillator to go from 0 to 2π is half the time required for the other to cross the same length (e.g. $\omega_2 = 0.5\ \omega_1$). To deal with this case, we can redefine Eq. $(xxiv)$ such that all possible modes of resonance, i.e. frequencies such that $\frac{\omega_1}{\omega_2} \simeq \frac{m}{n}$, are considered:

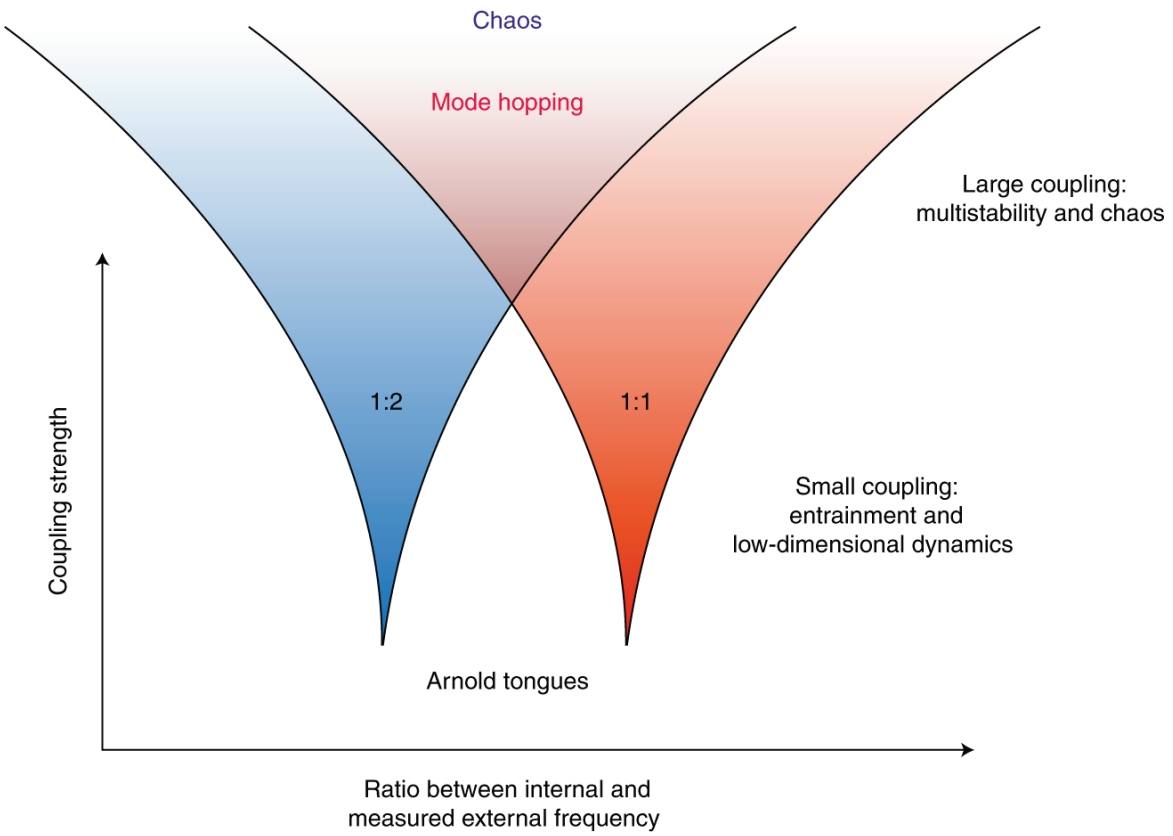
$$\Delta\phi = n\phi_1 - m\phi_2 \qquad (xxx)$$

Redoing the computations above, this yields the condition:

$$\epsilon q_{min} < n\omega_1 - m\omega_2 < \epsilon q_{\max} \qquad (xxxi)$$

In the case of noisy systems, the Langevin dynamics presented Eq. $(xxviii)$ still holds, with the exception that the potential has now several minima, corresponding to the different phase-lockings. Depending on the strength of the coupling, the noise can make the system jump from one phase-locking to another: this is called *mode-hopping*, and has notably been observed in the activity of the transcription factor nuclear factor κ B(NF κ B)[71].

This range of behaviours for deterministic and noisy systems is summarized in Introduction Figure 23, in which the synchronization states for different values of $n\omega_1 - m\omega_2$ and ϵ are plotted. This representation is called the Arnold tongues.



Introduction Figure 23: Representation of the Arnold tongues for an arbitrary oscillating system of two coupled oscillators. The 1:1 and 1:2 tongues indicate entrained states, where the numbers refer to the frequency of the external oscillator and the internal oscillator. In this way, 1:2 means that every time the external oscillator makes one rotation, the internal oscillator makes two rotations. Likewise, 1:1 means that the oscillators are synchronized in frequency. For higher coupling strength, if the system is noisy, mode-hopping can occur, in which the system jumps between multistable cycles. Above this, chaos emerges as the system can't adopt any stable phase-locking. Figure and caption taken from *Heltberg and Jensen, 2019* [72].

On this figure, synchronization occurs in the coloured zones. When there's no coupling ($\epsilon = 0$), synchronization can occur only when the ratio of the natural period is a purely rational number. However, as the coupling strength increases, the zones of synchronization expand, enabling two oscillators with quite different periods to synchronize nevertheless. The different tongues correspond to different modes of synchronization (i.e. phase-lockings). Depending on the coupling function(s) of the system, all the integer numbers fractions do not lead to synchronization, and for those who do, the synchronization area may be minimal. If the system is noisy and the coupling is strong enough, mode-hopping can occur, as the oscillators will jump from one phase-locked mode to another. If the noise keeps increasing, chaos emerges as the behaviour becomes completely unpredictable.

3.4.2. Synchronization of three or more oscillators

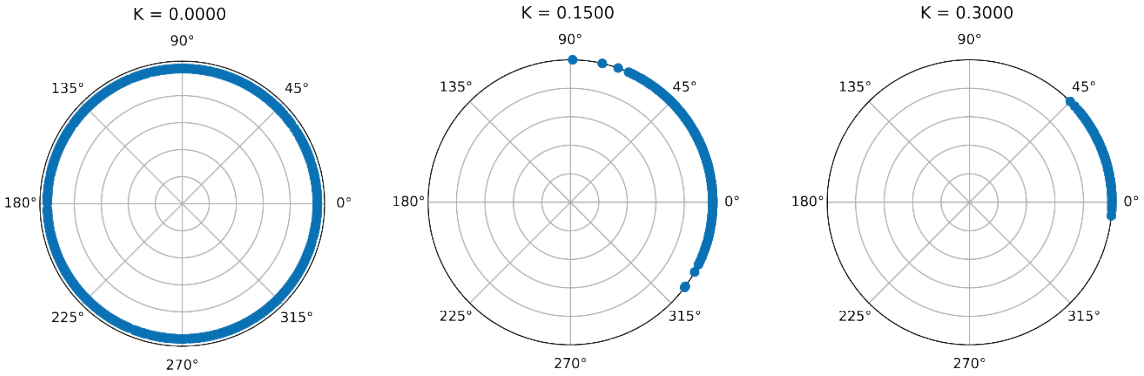
3.4.2.1. What synchronization means for large systems

In the general case, synchronization is harder to characterize for systems of more than two oscillators. It can be shown that a lattice of oscillators with nearest-neighbour coupling will exhibit either clusters of synchronization, wave-like behaviours, or no-synchronization, depending on the oscillators and coupling function(s) parametrization[73]. However, when oscillators are randomly interacting, things can quickly become intractable. Studies have been done on systems of three and four oscillators, yielding rather little guidance on the actual underlying theory[47].

Things are, however, more accessible when the interactions are homogeneous among oscillators. For instance, consider the following all-to-all coupled interacting system of oscillators, which is the simplest possible version of the famous Kuramoto model[74]:

$$\dot{\theta}_k = \omega_k + \frac{K}{N}\sum_{j=1}^{N} \sin(\theta_j - \theta_k), \text{ for } k = 1, \dots, N \qquad (xxxii)$$

Here, θ_j represents the phase of oscillator j, and K is the coupling strength. This system can easily be simulated, yielding various phase distribution at equilibrium depending on K, as shown in Introduction Figure 24 below.



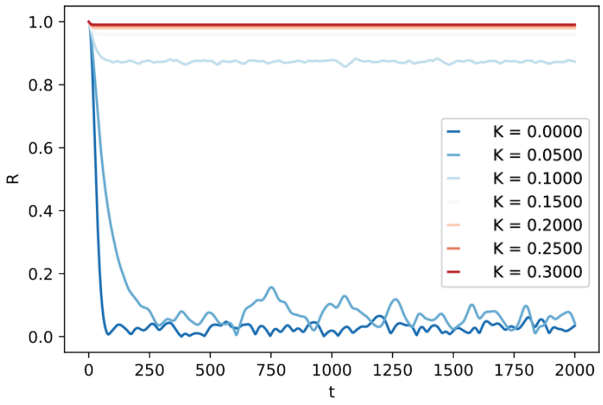
Introduction Figure 24: Representation on a polar plot of the evolution of the phase distribution of the Kuramoto model under all-to-all coupling with increasing coupling strength K (from left to right).

There exist several methods to quantify the degree of synchrony of such a system, but as the coupling is phase-attracting[18], the best is probably the synchronization index:

[18] It should be quite clear from the model that the coupling tends to make the oscillators' phase converge towards the same value, as $\sin(x)$ is an odd function.

$$R = \frac{1}{N}\left|\sum_j e^{i\theta_j}\right| \qquad (xxxiii)$$

We here express the magnitude R of the complex mean as an absolute value. If R is equal to 1, all oscillators have precisely the same phase. If R is 0, then the phases of the oscillators are evenly spread on the cycle (Introduction Figure 24, left). Plotting the evolution of R with the time, depending on the coupling strength, reveals an interesting pattern: when K exceeds a given threshold (about 0.1), all the oscillators tend to synchronize almost perfectly (Introduction Figure 25).



Introduction Figure 25: Evolution of the synchronization index with time, for different coupling strengths. The system exhibits two drastically different behaviours, as either the synchronization is quasi-perfect (top curves, in red), either non-existent (bottom curves, in blue).

This result can be proven, using tools from statistical physics, as the oscillating system shows a high similarity to spin systems under a mean-field model[32].

As a final remark, note that if synchronization exists but is not phase-attracting (i.e. oscillators tend to stick together, but with different phases), or if phases are not progressing linearly, phases should all be redefined using Eq. (v) to have the same value and same progression along a cycle. In any case, the synchronization index should always be computed after many periods, once the system has reached equilibrium. If $p:q$ modes of synchronization are observed, synchronization becomes intrinsically harder to quantify, but adapted versions of the synchronization index have also been developed[75]. Similarly, if a system is made of oscillators tending to synchronize by cluster (but with cluster having different phases, i.e. clique-like systems), a synchronization index has also been developed for that matter[76].

4. Fitting models to biological data

4.1. Introduction to modelling and inference

All models are wrong, but some are useful is a common saying in science, reflecting the idea that no model can truly capture the complexity of reality, but that sometimes, with some of them, one can still make relatively accurate predictions. There are many ways to construct a model, each being more adapted to a given type of problem. In practice, all these different methods are often intermingled, as a model is often judged more on its efficiency (the results it yields) than on the paradigm on which its structure was conceived.

The most intuitive method, and probably also the most accurate, tries to model the behaviour of the smallest components of a system to understand how it can generate complex behaviours. Such an approach is called *bottom-up*. Although they have demonstrated their usefulness in physics, these approaches have not been successful in biology. The reason is that biological systems are made from so many diverse parts (large number of degrees of freedom), which interact in such sophisticated manners, that it is usually impossible to retrace every single interaction[77]. And even if one could characterize all the parts and their interactions, the corresponding model would probably be intractable computationally. Things are progressively changing as, on the one side, experiments enable the collection of more data (high-throughput experimental techniques are becoming the norm in the omics era), and on the other side, computers are getting more and more powerful (to this day, Moore's law still applies). Nevertheless, overall, bottom-up methods are not well fitted for most biological problems.

The converse approach is called top-down, as it tries to understand a system by looking at the global picture rather than at the individual components. The objective, in the end, it to gain insight into the elementary sub-systems in a reverse engineering fashion. This approach usually works very well in biology, as many phenomena can be grasped without consideration of what happens at the smallest scale. For instance, physiology can easily be considered independently from genetics, although the former emerges from the latter.

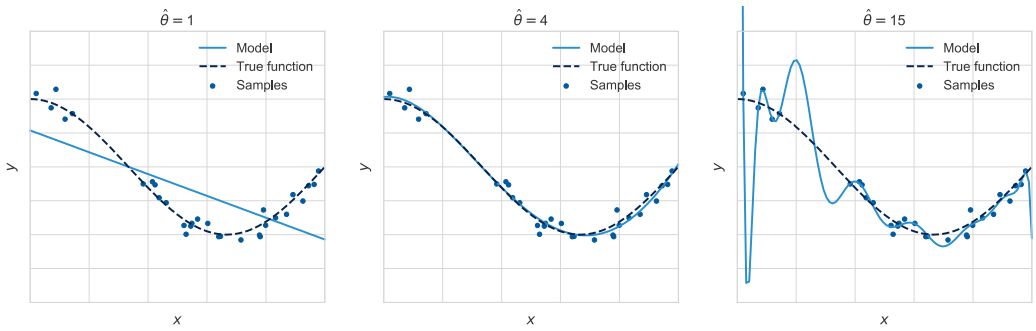
In conjunction with both bottom-up and top-down approaches, one can also try to identify signatures of universal mathematical principles such as bifurcations or synchronization. For instance, noisy oscillators tend to have very similar behaviours in physical, chemical and biological systems; although the underlying structure of their interacting components is very different. But, since it has been proven that there exist only a few ways by which oscillations can emerge, canonical models have been created (e.g. limit-cycle phase oscillators) that cover

a wide range of biological (or non-biological) systems, with no considerations for the qualitative nature of the small components of these systems.

Now, the previous approach works well given that one can accurately reproduce the behaviour of the system that is modelled. This is where the concept of *inference* arises. The principle of inference is to fit statistical models to data coming from experiments. This ensures that the model is not disconnected from reality, and can, if properly done, provide interpretable insights on the mechanisms at play or make interesting new predictions.

4.2. The problem of parsimony

A model with an infinite number of parameters can fit perfectly any dataset, but one will not learn anything from this model as it is no different from the data it is fitted to. Worse than that, such a model could actually bias the acquired knowledge due to the noise in the data. Conversely, if we were to choose a very simple model, for example, one described by just one parameter, the model would be straightforward to understand, but essential aspects of the dynamics might be left out. Therefore, a compromise must be made between simplicity, interpretability and fitting quality: this is the problem of parsimony. This compromise is illustrated in Introduction Figure 26, where some noisy samples are fitted with a polynomial of varying degree $\hat{\theta}$. For a low degree, the model is biased, but is easy to interpret. For an intermediate degree, the model is still relatively easy to understand, and shows a good fit with the data. For a high degree, the fit is perfect, but the model shows a very high variance, making it much harder to interpret, and highly decreasing its capacity to generalize. In practice, on real biological data, it is often hard to decide which parametrization yields the best compromise between bias and variance.



Introduction Figure 26: Illustration of the bias-variance trade-off, in which a set of datapoints is iteratively fitted by polynomials of increasing degree $\hat{\theta}$. The simpler fit (**left**) corresponds to a highly interpretable but highly biased model: this is called *underfitting,* as this kind of model is considered not precise enough. The intermediate fit (**middle**) corresponds to a less interpretable model showing more variance, but which is also less biased. A more complex fit (**right**) corresponds to a hardly interpretable model, showing a very high variance, but explaining almost entirely the given data: this is called *overfitting*, as this kind of model is unable to generalize to new samples. Figure created by adapting the code from the scikit-learn website[78].

Probably the most popular method used to choose the appropriate level of complexity for a given model was formulated by Akaike[79]. Combining Bayesian inference with information theory, Akaike managed to quantify how much information a model captures from a given dataset. In addition, he explicitly computed how much additional information would be gained when adding more parameters to the model. Combining these two quantities, one can determine a parsimonious model, where the model captures a high degree of information[80].

It is very often considered that a model with a large number of parameters is unfavourable. Von Neumann himself would even have formulated the problem this way: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk". However, this is not always true, for several reasons. First, the actual data that is modelled can be intrinsically complex, therefore requiring a large number of parameters. Then, not every model ought to be understood. In particular, machine-learning models, especially neural networks, tend to use a vast number of parameters, but this doesn't prevent them from making accurate predictions[19]. Finally, some parameters may explain transient dynamics that are ignored, or parts of the system that are simply not measured.

4.3. White and black box approaches

Along with the problem of bottom-up *vs* top-down and parsimony comes the issue of the type of parametric modelling approach one wants to take. Indeed, given some data, the modelling of the system under study can be done using either first principles or blind approaches. In the first case, known as *white-box* modelling, one must have an excellent knowledge of the physical (or other) mechanisms of the problem. If not, finding the proper model structure and complexity along with the appropriate parameters can be close to impossible. This type of modelling has, however, the advantage of being easily interpretable, and can be modified to consider the system's perturbations. Moreover, their behaviour can often be studied analytically[81].

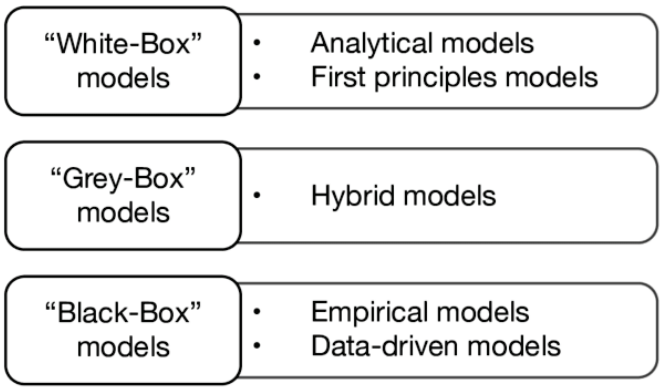
Often, however, one is not interested in the model itself, but rather in the prediction that can be made from it. Or, it can be that the system under study is so complex that white-box approaches are simply inapplicable. In such cases, using *black-box* modelling can be extremely useful. In practice, this means that one gives up on building a realistic structure for the model, as the corresponding parameters do not have any physical meaning. This also means that the parameters must be identified from scratch, with greedy optimization techniques. Analytical techniques are poorly applicable here. Machine learning techniques are

[19] However, it is true that neural networks suffer problems of overfitting, and it also true that they're often used as "black boxes". Cf. Section 4.3 below.

usually considered to be black-boxes, and, among them, neural networks are probably the most archetypical black-box models in existence. This is because this range of methods usually has several thousand parameters, whose values can significantly vary from one optimization to another, even on the same dataset. However, they tend to yield excellent results when properly trained, often much better than white-box approaches[82].

Finally, there exists one last range of modelling approaches known as *grey-box* or semi-physical modelling [83]. In these models, the parts of the system that are well known and understood have a realistic physical structure, while those that are not are modelled with ad-hoc approaches having a lot of free parameters. One example is the Monod saturation model for microbial growth[84]. This model is very similar to the Michaelis-Menten equations, but is considered more empirical. It links substrate concentration and growth rate with a simple scalar function of one parameter. This is physically wrong, but is efficient in practice and can be considered legitimate as one doesn't need to go into the details of chemical binding to model bacterial growth.

This model classification into black, white and grey boxes is summarized Introduction Figure 27 below.



Introduction Figure 27: A self-contained summary of the different modelling approaches. Taken from *Chambers, 2017*[85].

4.4. Inference as a mathematical problem

4.4.1. Likelihood

The problem of inference is always framed as a problem of likelihood maximization. The likelihood of a model describes how well it can explain the observed data. More precisely, the likelihood refers to the probability, given a model and a corresponding set of parameters, of obtaining a particular set of data. As a probability, likelihood should always be normalized, and, as a concept, it should always go along with a specific model that could have generated

the observed data. Calling the likelihood L, the model (with the corresponding set of parameters) M and the data D, this can be written mathematically as:

$$L(M|D) = p(D|M) \qquad (xxxiv)$$

In practice, the objective will be to find the set of parameters that yields the highest probability of the data, that is, the highest likelihood. The corresponding estimates for the parameters are then called maximum likelihood (ML) estimates. One problem of ML estimates is that they can be biased, especially when the sample size is small. Most of the time, this can be corrected *a posteriori* with regularization techniques (provided that the bias is known).

4.4.2. Parsimony

Another problem of ML estimates is that they don't explicitly account for the number of parameters in the model. Yet, the parsimony of a model is a complex problem, as explained in Section 4.2. Furthermore, one often wants to compare several competing models to decipher which one is the best. To this end, the Akaike Information Criterion (AIC) can be useful. Given a model with k parameters and likelihood L, the AIC can be computed as follow:

$$AIC(k, L) \;=\; 2k - 2\ln(L) \qquad (xxxv)$$

Eq. $(xxxv)$ can be intuitively understood as a trade-off between the likelihood of the model and the number of parameters used to fit the data. In sum, AIC can help to identify the model that describes patterns in the data the best, with the fewest number of parameters. In practice, as such, AIC is only valid for relatively large sample sizes compared to the number of parameters ($n/k > 40$, n being the sample size). If needed, a small sample size correction can be employed[80].

Similarly, an alternative criterium can sometimes be used: the Bayesian Information Criterium (BIC)[86]. The philosophy behind the BIC is very different as, when comparing several candidate models, it assumes that one of them is actually the true model, while the AIC simply tries to select the best model, knowing that it remains intrinsically wrong and describes an unknown, high dimensional reality[87].

4.4.3. Bayes formula

Likelihood computation can be approached using either frequentist or Bayesian inference. There's a vast literature regarding which one performs better in which setting, and

why one should be used over the other[88]. As the three studies presented in this thesis mainly use Bayesianism[20], I will only introduce the latter.

In a Bayesian framework, there are four quantities of interest:

- The prior probability of the parameters, $p(M)$, which is the probability one *thinks* that the parameters have, before fitting the data. This corresponds to prior knowledge about the probability that any given hypothesis is true, and this always remains a belief.
- The posterior probability of the parameters, $p(M|D)$, that is, the probability of the parameters given the data.
- The probability of the data given the parameters, i.e. the likelihood $p(D|M)$.
- The probability of the data, $p(D)$, that is, the expected probability of the data integrated over the prior distributions of the parameters.

All these quantities are linked through Bayes' theorem:

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)} \qquad (xxxvi)$$

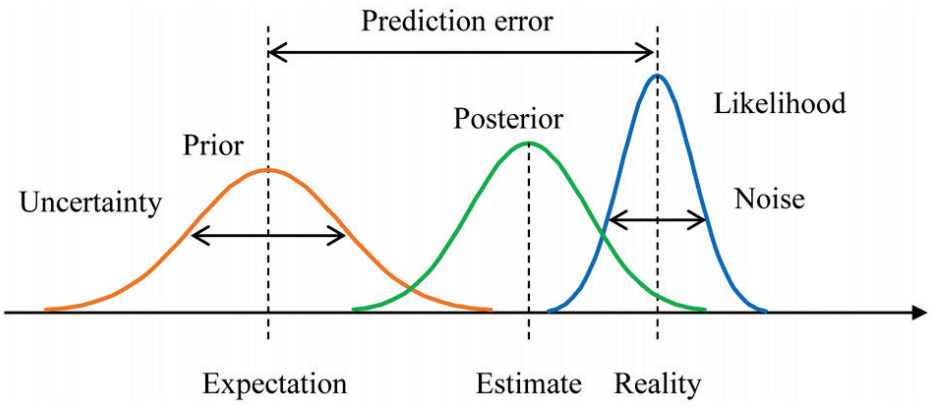
In practice, the probability of the data $p(D)$ is considered as a normalization factor, and the objective is simply to find the set of parameters M such that $p(M|D)$ is maximum:

$$p(M|D) \propto p(D|M)p(M) \qquad (xxxvii)$$

The general procedure is illustrated in Introduction Figure 28: one chooses a prior distribution $p(M)$ for the model parameters, with an uncertainty inversely proportional to the confidence attributed to it. This distribution is updated according to the corresponding probability of the data $p(D|M)$ (for the given model), yielding the final parameter distribution (after normalization): $p(M|D)$.

Note that the prior probability of the parameters must still be explicitly quantified. Often, one uses a non-informative prior (e.g. a uniform distribution) in order not to bias the posterior with uncertain belief.

[20] Although some standard statistical testing, which belong to the frequentist approaches, are also performed in these studies.



Introduction Figure 28: Cartoon illustration of Bayesian inference. The prior parameters distribution (in orange) is updated according to the data likelihood (in blue), to yield the posterior probability for the parameters (green). Figure taken from [89].

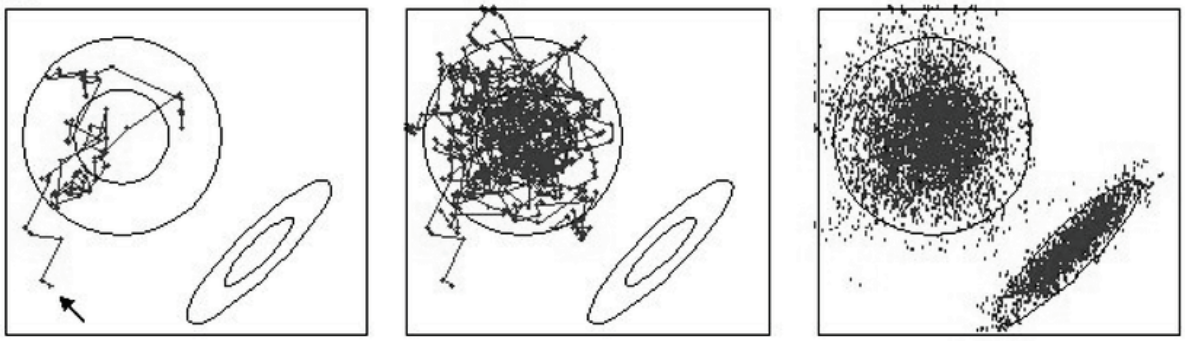
4.4.4. Optimization

In many cases, $p(M|D)$ can't be computed analytically. And even when it can be, it is often at the cost of approximations, and use of simplified distributions (e.g. Gaussians) without much supporting evidence. For simple problems, it is, however, easy to compute $p(M|D)$ numerically: continuous distributions can be approximated by random samples, and all the possible sets of parameters are successively tested, until one can rebuild a reasonable estimate of the posterior.

However, it is often the case that the computation of $p(D|M)$ involves hardly tractable sums or integrals, and that explicitly computing the posterior is simply not feasible. In this case, many alternative methods exist to sample the posterior in an unbiased and efficient way, the most famous one probably being Markov Chain Monte-Carlo (MCMC).

The principle of MCMC techniques is to use a chain of calculations to sample the posterior distribution (Introduction Figure 29). In practice, one builds a Markov chain that has the property of having the same equilibrium distribution as the posterior distribution of the model whose parameters are being optimized. Next, using a random simulation on this chain, one can directly sample from the posterior distribution of the model. When enough samples are collected, one can get a faithful reconstruction of the posterior, with far fewer computations needed than if an extensive search was done on the whole set of parameters[90].

However, often, one is not interested in the whole posterior distribution, but is simply looking for a realistic value for the parameters of the model. Such parameter estimates could be $\underset{M}{\operatorname{argmax}}\, p(D|M)$, or possibly $\mathbb{E}[p(M|D)]_M$. In this case, an extremely vast selection of optimization methods has been developed, each of them more or less specific to a given type of problem.



Introduction Figure 29: Illustration of MCMC principle, in which a random walk on a Markov chain progressively approximates a given posterior distribution (from left to right, number of iterations increasing). Taken from *Delsuc & Douzery, 2004*[91].

The difficulty is then to find which method is the most appropriate. Expectation-Maximization is one of these methods, used in the project *Low-dimensional dynamics of two coupled biological oscillators* and introduced Chapter 1, Section 1.2.3.2. L-BFGS is another, used in the project *RNA velocity-based inference of cell-cycle properties using single-cells* and introduced in Chapter 2, Section 1.2.2. Often, the model must be simplified as optimization is simply not tractable with conventional methods.

5. Concrete application

In this thesis, four concrete studies of noisy biological oscillators are presented, each of them building on the theory introduced above. We introduce them hereafter.

5.1. Low-dimensional dynamics of two coupled oscillators

This interdisciplinary physics project constitutes Chapter 1 of this thesis. It builds upon previous work from the Naef Lab aimed at understanding how two seemingly unrelated biological oscillators, the cell-cycle and the circadian clock, turn out to robustly synchronize under various environmental conditions. The cell-cycle and circadian clock system is very attractive for physicists and quantitative biologists since it's both sufficiently simple, i.e. "lives" on a low dimensional manifold, yet complex enough to exhibit universal dynamics.

In this project, the main objective is to understand the extent to which the cell-cycle influences the circadian clock, quantitatively. From there, one can answer many other questions about the system dynamics: what kind of synchronization is observed in the system? How robust is the synchronization? What are the biological implications of such dynamics?

To answer the questions above, we decided to use data-driven model reconstruction and nonlinear dynamics analysis to investigate the low-dimensional behaviour of the coupled

oscillators, in both mouse and human cells. We further studied the dynamics in physiological conditions and analyzed the influence of the cell-cycle on the circadian oscillator in tissues *in vivo*.

5.2. Space-time logic of liver gene expression at sublobular scale

This collaborative study with the Itzkovitz lab (Weizmann Institute, Israel) is presented in Chapter 2 of this thesis. It combines single-cell analysis of liver gene expression with temporal sampling and circadian rhythms to provide the first exhaustive analysis of mouse liver gene expression with both spatial and temporal resolution. Recently, the Itzkovitz lab combined single-cell RNA-sequencing of dissociated hepatocytes and single-molecule RNA fluorescence in situ hybridization to reconstruct spatial mRNA expression profiles along the liver central-portal axis. This revealed an unexpected breadth of spatial heterogeneity in mRNA expression that coincides with an intricate organization of spatially non-uniform liver functions. In parallel, the Naef lab showed how both the circadian clock and the feeding fasting cycles pervasively drive rhythms of gene expression in bulk, impacting key sectors of liver physiology.

We here decided to fill a critical knowledge gap by analyzing both spatial and temporal dimensions simultaneously. The main objective is here to understand how liver gene expression changes depending on both circadian time and lobule position along the central-portal axis. From there, one can also try to understand how liver functions are impacted by this spatiotemporal regulation, and if that makes sense from a physiological perspective.

To this end, we leveraged state-of-the-art single-cell, computational techniques and statistical analysis to report how spatial and temporal regulatory programs interact on the levels of individual genes and liver functions.

5.3. RNA velocity-based inference of cell-cycle properties using single-cells

In this collaborative study with La Manno's lab (EPFL, Lausanne), presented Chapter 3 of this thesis, we try to accurately predict the time evolution of single cells in the context of proliferative progress. More specifically, we take an interest in the cell-cycle, as it is one of the main drivers of cell-to-cell heterogeneity in gene expression in an otherwise homogeneous cell population.

The main objective is here to infer the cell-cycle progression state from a snapshot of the cell transcriptional state. From there, one can better understand how and why cell-cycle regulation differs among different cell-types, and potentially apply the same method to other differentiation processes.

This problem is not completely new, but although various methods have been developed to characterize cell-cycle progression, they usually rely on a few known markers, and suffer technical batch effects. Single-cell RNA sequencing approaches are also limited as they only provide a static snapshot of the cell expression states.

Here, we decided to take advantage of the high identifiability of the cell-cycle as a periodic 1D manifold in expression space to develop a new, cell-consistent version, of RNA velocity, to directly estimate the cell-cycle speed in any given cell. We validate our methods on several datasets, including various cell-types.

5.4. Phase inference

This technical review, presented Annexe A of this thesis, was aborted, mainly due to the concomitant publishing of similar reviews, but still contains reasonably exhaustive material. In there, I try to make a thorough summary of the available methods used for phase inference in the context of temporal signal processing. Phase inference is a fundamental problem as it is present in widespread areas of science (e.g. geology, neurosciences, biochemistry), but is often approached with inappropriate methods.

Four different classes of methods are presented, each having advantages and drawback, depending on the data under study. This comprehends linear and piecewise linear interpolation, Hilbert transform, smoothing approaches (e.g. Kalman filter) and Hidden Markov Models.

This whole project is available as an open-source GitHub repository.

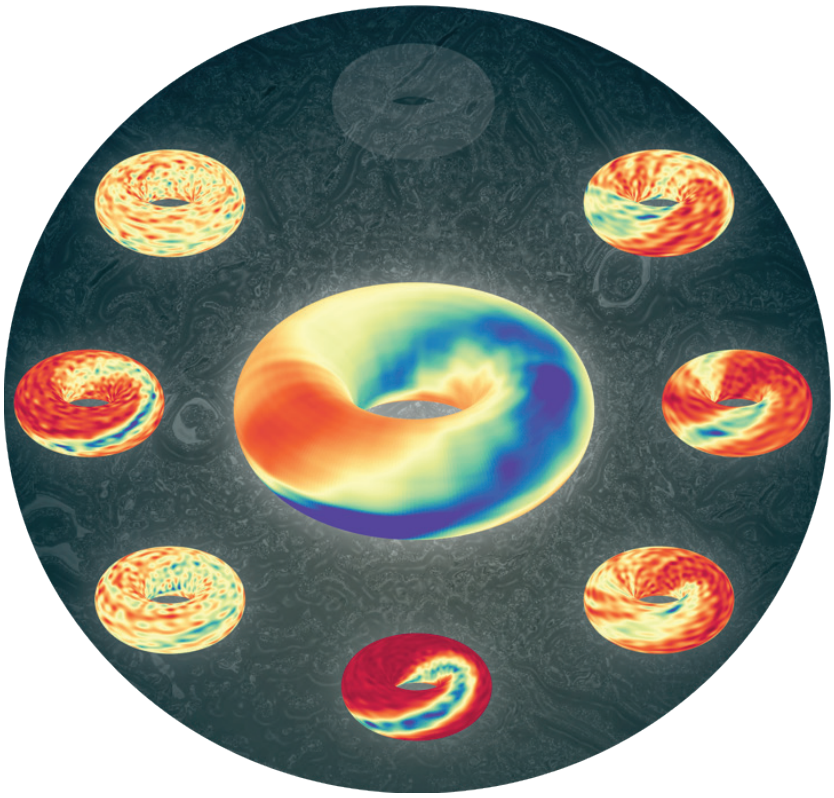
Chapter 1: Low-dimensional Dynamics of Two Coupled Biological Oscillators

This work was published in August 2019 in *Nature Physics* (vol. 15, no 10, p. 1086-1094), by **C. Droin, E. Paquet** and F. Naef (first authors in bold).

Contributions

In all the tasks listed below, my work is always under the supervision of F. Naef.

All the authors designed the study concept. I developed and implemented the whole computational framework (EM-HMM), and used it on the data coming from [45], which was cleaned beforehand by E. Paquet. I studied the system dynamics and model predictions (Figure 2-3-4). E. Paquet performed the supplementary experiments of the papers and analyzed the results (Figure 5-6). All the authors participated in the result interpretation and to the writing of the article, although I was mainly responsible for the parts linked to Figure 1-4 and E. Paquet for the parts linked to Figure 5-6. The GitHub code provided is mine.



Artwork figure 2: Artistic representation of the different phase-lockings (peripheral toruses) observed in the coupled system linking the cell-cycle and circadian clock dynamics. The coupling function is represented on the central torus. The background is a brightfield microscopy capture of mice fibroblasts.

1. Project introduction

1.1. Motivation and aims

In order to generate complex adaptive behaviour, many biological oscillators are coupled to external driving forces (e.g. light cues or food intake can reset the circadian clock[92]), or to other oscillators (e.g. circadian gating of the mitosis in plants, zebrafish, and cyanobacteria[93]). Although the mechanisms behind these couplings have been under investigation since the 1950s[25], the knowledge of the molecular interactions involved in the process remains still limited. Similarly, the quantitative aspects of the coupling are still not well understood for a large proportion of the biological systems.

This lack of research could partly be explained by the absence of efficient analytical tools. Things have changed as, recently, the emergence of reporter systems, as well as with the progress of microscopy techniques and image analysis, enabled the measurement of temporal traces as proxies for tracking oscillating systems. For instance, this includes circadian oscillation[94], or metabolic oscillations[95].

However, unravelling how the systems are coupled from noisy temporal traces is still a hard problem. Indeed, this implies modelling the traces as oscillatory signals and finding how the underlying physical phases influence each other. Classical signal analysis methods, such as Fourier or Hilbert transforms are not well adapted since the microscopy traces are usually noisy, with variation in amplitude, signal background and phase delays. Cross-correlation techniques would only inform if the system is synchronized, but not how. Moreover, these methods are essentially top-down approaches, which don't tell us much about the fundamental parameter changes in the model underlying the system. Finding a simple way to analyze the behaviour of a system of one or several coupled oscillators is, therefore, a problem of great interest, with a broad scope of biological applications.

In this project, we wanted to develop a computational method to study the behaviour of a system of two noisy biological oscillators. Finding the instantaneous phase of the oscillators, as well as unravelling how they interact were among the questions of prime importance which we wished to answer. In parallel, we wanted to get a better understanding of the corresponding dynamical system. For instance: what are the corresponding Arnold tongues? Can we predict new phase-lockings with our model?

In practice, we decided to work in the continuity of the project developed by J. Bieler[45]. As such, we used a statistical framework of inference based on a Hidden Markov Model (HMM), to study the problem of dynamical coupling between the circadian clock and the

cell cycle in mammalian cells. This system was very relevant as the phase combination for which the cell-cycle speeds up or slows down the circadian clock was still unknown, although the question had been approached a few times in the past[44], [45].

1.2. Background

1.2.1. Two interacting biochemical oscillators

1.2.1.1. The mammalian circadian clock

During Evolution, organisms learnt to adapt to the predictable daily variations of environmental conditions going with the Earth's rotation. Cyanobacteria were among the first to acquire a biological clock, later followed by plants, fungi and animals. The circadian clock gives an estimation of time and allows for coordination of physiology and metabolism in anticipation of recurring changes. It is involved in the functioning of several fundamental processes: digestion, sleep/wake behaviour, body temperature, cell-cycle control, metabolism etc.[8]

The molecular study of the mammalian circadian clock started at the beginning of the 1970s with the discovery of the central pacemaker: the suprachiasmatic nuclei of the hypothalamus (SCN)[97], [98]. The SCN is made of several interlocked feedback loops relying on a few clock genes (Background Figure 1.1, left). The main loop involves the genes CLOCK and BMAL1, which form a dimer to activate the transcription of the period genes (*Per1, Per2, Per3*) and cryptochrome genes (*Cry1, Cry2*). In turn, PER:CRY proteins will repress their own transcription *via* direct interaction with CLOCK:BMAL1. This first loop is then tuned by two secondary loops, in which the REV-ERBα,β and RORα,β,γ proteins operate respectively a negative and a positive feedback on the transcription of Bmal1. Post-translational processes are also involved in the regulation of this genetic network. An interesting property of the molecular clockwork is that it can act completely autonomously, and even in the absence of resetting cues, it still oscillates with an endogenous period close to a day. In the presence of environmental parameters (e.g. light, food), the synchronization to a 24h period becomes sharper. The SCN can then synchronize the entire organism *via* humoral cues[99]. This synchronization affects the local oscillators that operate in the cells of most organs and tissues. Each individual cell can thus be regarded as possessing its own circadian oscillator[100], [101].

1.2.1.2. The eukaryotic cell-cycle

For an organism to grow and maintain homeostasis, its cells must be capable of dividing. This is done through a process called the cell division cycle. The length of the cell-cycle is extremely variable from cell type to cell type, and even from cell to cell. For many

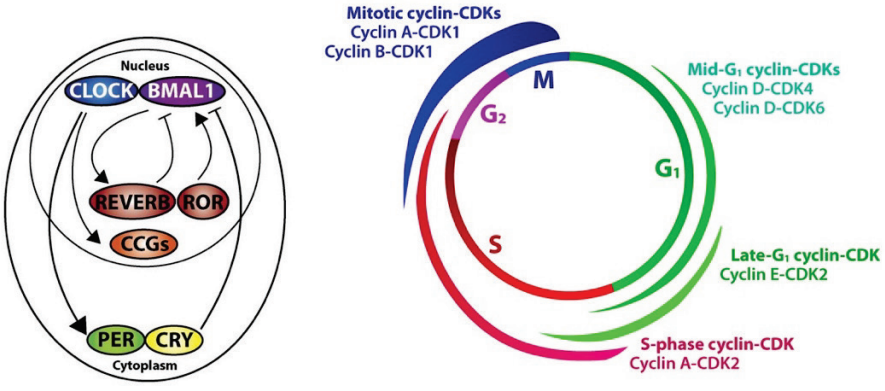
actively dividing mammalian cells, it is of the order of 24h (e.g. fibroblasts), but for fast-cycling cells, like the ones that line the intestine, this can be down to half this time[102]. For embryonic cells, some growth phases can be skipped, leading to cell-cycle with periods of the order of a few hours[103].

During the cell-cycle, two critical events happen: the synthesis phase (a.k.a. S phase), which is particularly important since the cell must take the time to replicate the several million bases constituting its DNA, and the division of the cell (a.k.a. M phase, for mitosis). Each of these phases is preceded by a growth phase: G1 before synthesis, and G2 before mitosis.

During the G1 phase, the cell grows in size and synthesizes mRNA and proteins that are required for DNA synthesis. This phase is crucial in the cell cycle because this is when a cell decides to commit to division or to leave the cycle and go to a quiescent state. This decision is made according to the success of passing the G1/S checkpoint. During this checkpoint, the cell checks that it has grown enough, that its DNA is not damaged and that the environment is rich enough in nutrients[103].

During the G2 phase, the cell will continue to grow. The G2 checkpoint control mechanism ensures that everything is ready to enter the mitosis: no UV radiation, oxidative stress, or DNA intercalating agents are present, and the DNA replication went as planned. If everything is alright, the cell enters in M phase and divides.

After division, the two daughter cells can either continue in the cycle (i.e. go to G1 again) or stay in a quiescent state: the G0 phase. In the second case, they may later resume cell division, depending on environmental parameters such as growth factors[104].



Background Figure 1.1: The cell-cycle and the mammalian circadian clock. Left: Simplified representation of the mammalian circadian clock, including the two main transcription-translation negative feedback loops, adapted from the *Chaix & al.* review[106]. **Right:** simplified representation of the eukaryotic cell-cycle, with the associated cyclin-cdk regulating complexes, adapted from the article by *Levi*[107].

At the molecular level, the progression of the cell cycle relies on the transient and sequential activation of cyclin-dependent kinases (abbreviated as CDKs), which form complexes with proteins called cyclins. Successively, the cell-cycle depends on cyclin D and CDK4–6 (G1 phase), cyclin E and CDK2 (G1/S transition), Cyclin A and CDK2 (S phase), Cyclin A and CDK1 (S/G2 transition), and cyclin B and CDK1 (M phase)[105] (see Background Figure 1.1, right, for a cartoon of the different phases of the cell-cycles, as well as with the cyclin-CDK complexes necessary to pass the different phases).

The activity of these cyclin-CDK complexes precisely modulates each phase of the cell-cycle, especially around critical checkpoints. CDK inhibitors (CKI–P16, P27, P21) or phosphorylation by the kinase WEE1 can inhibit the activity of targeted cyclin-CDK complexes across the cycle. Conversely, phosphatases such as CDC25A,B,C can activate the enzymatic complexes. Some of these activators/inhibitors are targeted by proteins involved in the repair of the DNA, thus blocking the cycle if the DNA is damaged or currently under repair. For instance, a double-strand DNA break activates the ataxia telangiectasia mutated (ATM) and checkpoint kinase 2 (CHK2) proteins, while a single-strand break or a replication error activate ataxia-telangiectasia related (ATR) and checkpoint kinase 1 (CHK1) proteins. These complexes cause a cell cycle arrest by indirect induction of CKI[96].

1.2.1.3. Coupling between the cell-cycle and the circadian clock

One interesting property of both the cell cycle and the circadian clocks is that they display periodic phases of activation and repression[108]. Since they can both be considered as autonomous biological oscillators, and since they coexist in the same dividing cells, one may wonder if they interact and whether this can have dynamical consequences on the two systems. This problem, termed as coupling between the cell-cycle and the circadian clock, is still an open question. Interestingly, pioneer studies have proposed that interactions occur in both directions. Hereafter are presented some of the results concerning the mechanisms in play at both the mechanistic molecular level, as well as the quantitative level.

1.2.1.3.1. Molecular mechanisms

1.2.1.3.1.1. Influence of the circadian clock on the cell-cycle

The influence of the circadian clock on the cell cycle is supposed to occur either by transcriptional control or direct protein-protein interaction. In the G1 cell-cycle phase, the cyclin-dependent kinase inhibitor (CKI) P21 is transcriptionally regulated by REV-ERB-α,β and ROR-α,β,γ [109]. At the G1/S transition, NONO regulates the p16-Ink4A checkpoint gene in a PER-dependent fashion[110]. The G2/M transition is controlled by the transcription of the WEE1 kinase, whose transcription itself is controlled by the CLOCK:BMAL1

dimer[111]. Looking at the post-translational level, CRY modulates the G1/S transition checkpoint through CHK1/ATR by interacting with TIM in a time-dependent manner. PER and TIM also regulate the G2/M transition via interactions with CHK2-ATM[112], [113]. Oncogenes, Cyclins and the tumour suppressor p53 are also known as clock-controlled cell-cycle regulators.

1.2.1.3.1.2. Influence of the cell-cycle on the circadian clock

The influence of the cell-cycle on the circadian clock also seems to occur at several layers. The most obvious would be the transcriptional shutdown occurring around mitosis[114], which is supposed to alter the circadian feedback loops. It has also been shown that DNA damage can advance the circadian phase in a dose and time-dependent manner, possibly through the involvement of PER and TIM proteins[115]. The tumour suppressors P53, as well with the promyelocytic leukaemia proteins could also influence the circadian function: Per2 transcription is repressed by P53, which in turn prevents the binding of the circadian complex CLOCK:BMAL1. After translation, PML physically interacts with PER2, and promotes its nuclear localization. These molecular connections are supposed to alter global circadian behaviour [116].

1.2.1.3.2. Quantitative characterization of the coupling

In many unicellular eukaryotes and cyanobacteria, the circadian system controls the timing of cell division. This originates from the experiments made on the cyanobacterium *Synechococcus Elongatus* as well as on the flagellate alga *Euglena Gracilis*, in which the molecular clock imposes a "gating" on the cell division at specific circadian phases[117], [118].

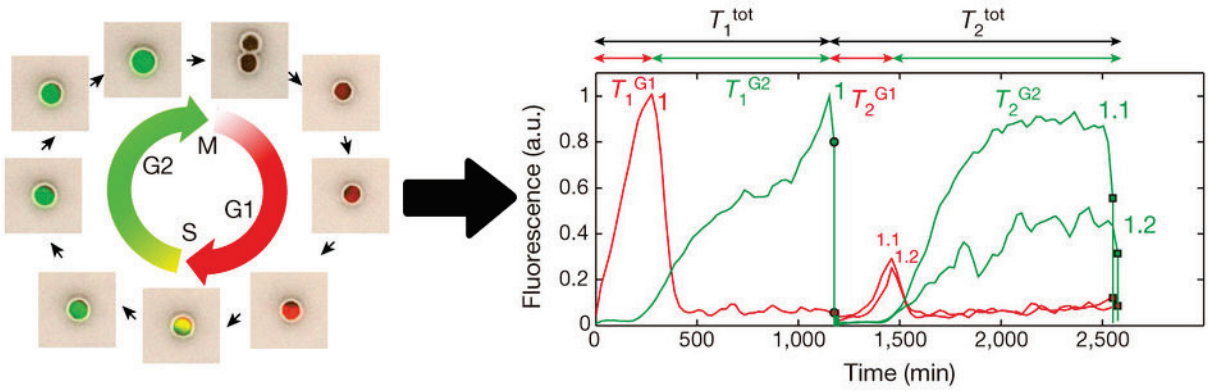
It could be tempting to extrapolate this gating phenomenon to pluricellular organisms, but the research made has only led to controversial evidence. While studies from the team of Nagoshi[100] (*in vitro*, with NIHT3T fibroblasts) and Matsuo[111](*in vivo*, in the mouse liver) reported a gating of the circadian clock on the time of mitosis, more recent studies did not report such a gating[44], [119]. Other studies suggest control of the cell-cycle by the clock, but with no direct evidence, on human mucosa and skin[120], mouse bone marrow[121], and many other cell lines (see review[93]).

The past work of our team[45] (NIH3T3 specific), indicated that, conversely, the circadian clock was under the influence of the cell-cycle. This is in agreement with what the findings of the Rand lab[44], apart from the fact that they concluded that the influence was bidirectional (although weak from the clock to the cell-cycle), while our team found a seemingly very low, likely non-existent, influence of the clock on the cell-cycle.

Therefore, there is still no clear answer to this question of mutual coupling. If it seems likely that the influence of the clock on the cell-cycle seems weak (at least in studies *in vitro*), the converse interaction remains unclear. Both our team and Feillet team investigated this problem, but both used imprecise or unreliable methods: in [44], *Feillet & al.* assumed that the circadian phase increased linearly between two peaks, therefore making impossible to infer a punctual coupling. In [45], our team, although not assuming a linear phase between the circadian peaks, only used the peaks times to infer the coupling functions (in the framework of a parametric model), leading to an incomplete inference.

1.2.2. Data collection

Many biological oscillators can be observed at the single-cell level. To know how they behave and interact, molecular tools have been developed to track their activity in real-time, at both the population and the single-cell level. For instance, in the case of the circadian clock, fluorescent and/or luminescent reporters enable to follow the evolution of the core-clock protein or mRNA concentrations[122]. Similarly, for the cell-cycle, morphological criteria (e.g. mitosis, nucleus size) can bring information on the oscillator phase, but a two-colour FUCCI cell-cycle sensor and its derivatives remain the most precise methods of phase tracking up-to-date[123]. Other analogous reporters exist for various types of oscillators: sleep-wake activity patterns (circadian clock), dissolved oxygen concentration (metabolic cycle in yeast), EEG (neural waves), etc.



Background Figure 1.2: Example of pipeline used to track the cell-cycle oscillator with the two-colours FUCCI sensor. In addition to the schematic description of the FUCCI markers used for visualizing the cell-cycle progression, time-lapse images of the dividing cells are given, showing overlays of phase-contrast green and red fluorescence images. From these images, plots of the FUCCI fluorescence traces can be extracted, and the period of the green and red reporters can be computed. Adapted from the work of *Sandler & al.*[124].

In most cases, fluorescent/luminescent reporters are used, since they enable to track the expression of any desired genes. Movies are then obtained from time-lapse microscopy. These movies are then segmented using image processing techniques, and the essential features (e.g.

fluorescence concentration of a given nucleus) are extracted; in our case, this leads to the production of raw temporal traces. Note that the segmentation can be challenging since the images can be overexposed for some cells (saturating the detector) or underexposed (making the cells hard to distinguish from the background). Besides, there is also significant cell-to-cell variability in the fluorescence activity.

A representation of the procedure used to track the cell-cycle phase using the two-colour FUCCI sensor is given Background Figure 1.2.

1.2.3. Inference

1.2.3.1. Challenges

Ideally, one would like to infer the oscillators behaviour directly from temporal traces. This is possible using an approach based on maximum likelihood, in which one starts from a simple and generic model, which is then parametrized in a way that explains best the data. To that end, three critical challenges must be faced:

- Modelling the oscillatory experimental signals (e.g. Background Figure 1.2, right)
- Estimating the parameters of our model. In the case of a system of coupled oscillators, this includes the coupling function(s).
- Estimating the instantaneous phases from noisy experimental signals.

The starting data consists of experimental signals (circadian reporters and cell-cycle reporters). To deal with the first challenge, one needs to build a model taking as input the system's phases, and giving as output the desired periodic signal. This may be difficult as the experimental signals are noisy, may have a variable background, variable peaks amplitude, and finally, the peak-to-peak intervals may be very inconstant from one trace to another, or even in the same trace.

Given a theoretical model for the likelihood of the measured data, one wishes to find a way to estimate a set of parameters $\boldsymbol{\Lambda}$ that maximizes the probability of the data D. The second challenge can thus be formulated as finding $\boldsymbol{\Lambda}^*$ such that:

$$\boldsymbol{\Lambda}^* = \mathrm{argmax}_{\boldsymbol{\Lambda}} p(\mathrm{D}|\boldsymbol{\Lambda}) = \mathrm{argmax}_{\boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\Lambda}|\mathrm{D}) \qquad (i)$$

In this equation, $\mathcal{L}$ is the data likelihood. Depending on the number of parameters to optimize, as well as the size of the data D, the optimization process may be hardly tractable with naive methods.

Finally, having the model and the parameters, one must estimate the underlying phase of the system from the signal. The data being very noisy, a Bayesian approach seems more adapted; the third challenge would, therefore, be to compute the probability distribution of the phase $\boldsymbol{\Phi}$ given the data D and the parameters $\boldsymbol{\Lambda}$: $p(\boldsymbol{\Phi}|D,\boldsymbol{\Lambda})$.

To solve the challenge 1, we decided to choose a parsimonious solution: the phase progression is converted into periodic oscillations using a waveform function. Variations in amplitude, vertical shifts and noise are tuned using simple multiplicative or additive terms.

To solve challenge 2, one needs to find the most appropriate optimization process for the system. Given that we have more than a thousand parameters due to the coupling parametrization, naive methods such as grid-search or gradient descents are hardly tractable. We decided to choose an Expectation-Maximization (EM) approach as this algorithm handles a large number of parameters and is guaranteed to converge. In addition, it can easily be extended to include constraints to regularize the obtained solution. The principle of the EM algorithm is developed in Section 1.2.3.2.

To solve challenge 3, one needs to find a statistical method of inference outputting the probability distribution of the underlying phases given the experimental signals. We opted for a well-known algorithm in the field of dynamic Bayesian networks: the Hidden Markov Model (HMM). The principle of the HMM is developed in Section 1.2.3.3.

1.2.3.2. The Expectation-Maximization algorithm

The EM algorithm is a general method for finding the maximum-likelihood estimate of the parameters of a hidden or underlying distribution from a given data set. It was first presented in 1977 in the paper of *Dempster & al.*[125]. However, the explanations that follow are adapted from the work of *Jeff Bilmes*[126].

As explained in Section 1.2.3.1, the objective in this project is to find the set of parameters $\boldsymbol{\Lambda}$ which maximizes the probability of the data $D = \{\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_N\}$. Here, the d_i must be interpreted as data points drawn from a distribution $p(\mathrm{d}|\boldsymbol{\Lambda})$. The resulting likelihood for all samples is:

$$\mathcal{L}(\boldsymbol{\Lambda}|\mathrm{D}) = \prod_{i=1}^{N} p\,(\mathrm{d}_i|\boldsymbol{\Lambda}) \qquad (ii)$$

The problem is that we can't directly compute $\mathcal{L}(\boldsymbol{\Lambda}|\mathrm{D})$ since the distribution $p(\mathrm{d}_i|\boldsymbol{\Lambda})$ is unknown. Indeed, with the inference framework that we use, the probability of observing the i^{th} experimental datapoint, given a set of parameters $\boldsymbol{\Lambda}$, actually depends on the distribution

of hidden phases, which itself depends on the set of parameters we choose. The problem must, therefore, be stated differently.

The solution to this problem is to introduce a latent variable $\boldsymbol{\Phi}$ (in our case, representing the hidden phases), such that the maximum likelihood estimate of the unknown parameters is determined by the marginal likelihood of the observed data:

$$\mathcal{L}(\boldsymbol{\Lambda}|\mathrm{D}) = \int_{\boldsymbol{\Phi}\in\boldsymbol{\Upsilon}} p(D, \boldsymbol{\Phi}|\boldsymbol{\Lambda}) d\boldsymbol{\Phi} \qquad (iii)$$

In this equation, $\boldsymbol{\Upsilon}$ is the domain of the phase values. The quantity $\mathcal{L}(\boldsymbol{\Lambda}|D)$ is often intractable (in our case, since $\boldsymbol{\Phi}$ is a sequence, the number of values it can take grows exponentially with the sequence length, making the exact calculation of the sum impossible). The EM enables to find the maximum likelihood estimate of the marginal likelihood by iteratively applying two steps.

The first step is to compute the expectation of this likelihood with respect to the hidden phases $\boldsymbol{\Phi}$, computed with the current set of parameters. The obtained expression, called Q, depends on a new set of parameters:

$$Q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') = E_{\boldsymbol{\Upsilon}}\big[\log\big(p(\mathrm{D}, \boldsymbol{\Phi}|\boldsymbol{\Lambda})\big)\big|\mathrm{D}, \boldsymbol{\Lambda}'\big] \qquad (iv)$$

Since $\mathrm{D}, \boldsymbol{\Lambda}'$ are here given, this expression can be developed:

$$Q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') = \int_{\boldsymbol{\Phi}\in\boldsymbol{\Upsilon}} \log\big(p(\mathrm{D}, \boldsymbol{\Phi}|\boldsymbol{\Lambda})\big) p(\boldsymbol{\Phi}|\mathrm{D}, \boldsymbol{\Lambda}') d\boldsymbol{\Phi} \qquad (v)$$

Note that in this equation, $p(\boldsymbol{\Phi}|D, \boldsymbol{\Lambda}')$ will be computed using a HMM (it's the posterior decoding). Evaluating the function Q is the first step of the EM algorithm, a.k.a. the expectation step. The second step is to maximize this expectation according to a new set of parameters:

$$\boldsymbol{\Lambda}'' = \mathrm{argmax}_{\boldsymbol{\Lambda}} Q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') \qquad (vi)$$

It has been shown that iteratively alternating between the expectation and maximization steps is guaranteed to increase the log-likelihood of the data, with a monotonous convergence[127]. Note that the algorithm can get stuck in local maxima and that it must, therefore, be run several times with different seeds.

We can sum up the inference in the following way:

1. Assume a set of parameters $\boldsymbol{\Lambda}'$ (for instance, the period of the oscillator, noise, etc.).

2. Run the HMM on the experimental points D, obtained from time-lapse imaging, to yield a given distribution of hidden phases $p(\boldsymbol{\Phi}|D,\boldsymbol{\Lambda})$.

3. According to this distribution of hidden phases, compute a better value of the parameters $\boldsymbol{\Lambda}$.

4. Repeat until convergence of the likelihood of the data.

As a final point on the EM algorithm, it is interesting to mention the fact that it can easily be extended to consider penalties. For instance, if we know that some parameters cannot take very high values, or if we need to keep some part of the parameter space smooth, we can penalize the likelihood with a function $J(\boldsymbol{\Lambda})$, according to new parameter λ:

$$\mathcal{L}(\boldsymbol{\Lambda}|\mathrm{D}) - \lambda J(\boldsymbol{\Lambda}) \qquad (vii)$$

In this case, it can readily be shown that one can run the same procedure and maximize the following expression:

$$Q(\boldsymbol{\Lambda},\boldsymbol{\Lambda}') - \lambda J(\boldsymbol{\Lambda}) \qquad (viii)$$

However, taking the derivative of J with respect to $\boldsymbol{\Lambda}$ can often lead to complex updates. But it has been proved that a one-step late update (that is, replace $J(\boldsymbol{\Lambda})$ by $J(\boldsymbol{\Lambda}')$) also converges[128].

1.2.3.3. Hidden Markov Model

The model representing the phase dynamics that we chose to study consists of a system of stochastic differential equations with variables evolving in a closed domain. This implies that the data likelihood computation can be expressed in the formalism of a Hidden Markov Model.

As an intuitive approach, a HMM is just like a classic Markov chain, except that each state of the chain can emit one or several observations whenever occupied. The observer only has access to the sequence of observations[21], and usually wishes to uncover which state produced which observation, and when. In our case, the hidden states correspond to the different possible values of the oscillator phase, and the observations are the temporal traces (e.g. fluorescence traces of a circadian reporter).

A formal definition of a HMM is given in the book of *Cappé & al.*[129]: "A HMM is a bivariate discrete-time process $\{X_t, O_t\}_{t\geq 1}$, where $\{X_t\}_{t\geq 1}$ is a Markov chain and, $\{O_t\}_{t\geq 1}$ is a

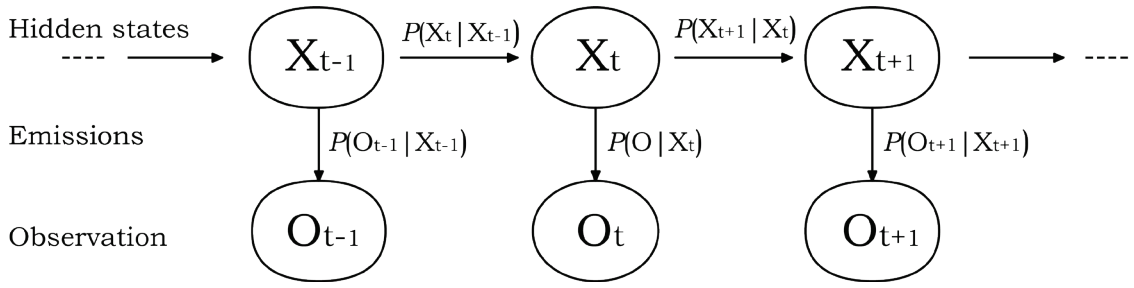
[21] That's why the states of the Markov chain are called "hidden states"

sequence of independent random variables conditional on $\{X_t\}_{t\geq 1}$ and such that the conditional distribution of O_t only depends on X_t."

As a reminder, a first-order Markov chain is a stochastic process for which the Markov property to the first order is verified, namely:

$$p(X_t|X_0,\ldots,X_{t-2},X_{t-1}) = p(X_t|X_{t-1}), \forall t \geq 1 \qquad (ix)$$

Therefore, the characterization of a HMM involves the specification of a transition and an emission probability distribution: $p_{tr}(X_t|X_{t-1})$ and $p_e(O_t|X_t)$. A schematic representation of how these probabilities rule the state transitions is given in Background Figure 1.3.



Background Figure 1.3: Schematic representation of the HMM for the transition between the states X_{t-1}, X_t, X_{t+1} and the corresponding emissions.

Moreover, since the Markov chain is a time-dependent process, we need an initial probability distribution $\boldsymbol{\pi} = p(X_1)$. Finally, since we're dealing with discrete states, we need to specify the state spaces for the hidden states and for the observations: $\mathcal{X} = \{x_1,\ldots,x_N\}$ and $\mathcal{O} = \{o_1,\ldots,o_M\}$ respectively (note that, to keep things simple, we here assume that the observations are one-dimensional). Using these discrete spaces, we can store the state transition and emission probabilities in resp. the matrices $\mathbf{A}$ and $\mathbf{E}$, such that $\mathbf{A}_{ij} = p_{tr}(X_{t+1} = x_j|X_t = x_i)$ and $\mathbf{E}_{ij} = p_e(O_t = o_j|X_t = x_i)$. A HMM can, therefore, be summarized as a stochastic quintuplet model $\boldsymbol{\Omega} = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{E}, \mathcal{X}, \mathcal{O}\}$.

HMMs are mainly used to solve three problems[130]:

1. Given a model $\boldsymbol{\Omega}$ and a sequence of observations $O = \{O_1,\ldots O_T\}$, find $p(O|\boldsymbol{\Omega})$. In other words, determine the likelihood of the observations given the model. This can be done using the forward algorithm.

2. Given a model $\boldsymbol{\Omega}$ and a sequence of observations O, find an optimal states sequence for the underlying Markov process. This can be done with the Viterbi algorithm.

3. Given an observation sequence O and the spaces $\mathcal{X}$ and $\mathcal{O}$, find A,E,$\boldsymbol{\pi}$ such that the likelihood of O is maximum. This can be viewed as training a model to best fit the observed data. This problem can be solved with the Baum-Welch algorithm.

Our research project involves the resolution of variants of these three problems: the first one is the parameter optimization, which is close to problem 3 above. However, in our case, A and E are generated using parameterized equations, meaning that the training must be done using a more abstract class of algorithms than the Baum-Welch algorithm: the EM algorithms (cf. Section 1.2.3.2). In addition, this training involves the computation of a likelihood function (cf. problem 1); this can be done using the forward algorithm. The last problem is a variant of the problem 2 above: we want to find the entire distribution of probabilities of the hidden states (not just the most likely sequence), in order to compute a coupling function using a completely Bayesian framework. Therefore, we don't use the Viterbi algorithm, but a more powerful (and greedier) algorithm: the forward-backward algorithm.

2. Published article

The article hereafter has been as little modified as possible. According to EPFL recommendations, the reference numbers are in the continuity of the thesis. Equation and Figure numbers are the same as in the published format (although the Chapter number is now indicated in the Figure captions).

2.1. Abstract

The circadian clock and the cell cycle are two biological oscillatory processes that coexist within individual cells. These two oscillators were found to interact, which can lead to their synchronization. Here, we develop a method to infer their coupling and non-linear dynamics from thousands of mouse and human single-cell microscopy traces. This coupling predicts multiple phase-locked states showing different degrees of robustness against molecular fluctuations inherent to cellular scale biological oscillators. Moreover, the phase-locked states were temperature-independent and evolutionarily conserved from mouse to human, hinting at a common underlying dynamical mechanism. Finally, we detected a signature of the coupled dynamics in a physiological context, where tissues with different proliferation states exhibited shifted circadian clock phases.

2.2. Introduction

The circadian clock and the cell cycle are two periodic processes that cohabit in many types of living cells. In single mammalian cells, circadian clocks consist of autonomous feedback loop oscillators ticking with an average period of about 24h[100], and controlling many downstream cellular processes[131]. In conditions of high proliferation such as those found in cultured cells or certain tissues, the cell-cycle progresses essentially continuously and can thus be abstracted as an oscillator with an average period matching the cell doubling time. Both processes fluctuate due to intra-cell molecular noise, as well as external fluctuations. While the precision of the circadian period is typically about 15% in fibroblast cells[100], the cell cycle can be more variable depending on the conditions and cell lines[132], [133]. Interestingly, previous work showed that the two cycles can mutually interact[100], which may then lead, as theory predicts, to synchronized dynamics[44], [45] and important physiological consequences such as cell-cycle synchrony during liver regeneration[111]. In tissue-culture cells, which are amenable to systematic microscopy analysis, it was found that the phase dynamics of two oscillators shows phase-locking[44], [45], defined by a rational rotation number $p:q$ such that p cycles of one oscillator are completed while the other completes q.

Concerning the nature of those interactions, the influence of the circadian clock on cell-cycle progression and division timing has been shown in several systems[110], [111], [134]–[137]. In contrast, we showed in mouse fibroblasts that the cell cycle strongly influences the circadian oscillator[45], which was also investigated theoretically and linked with DNA replication in bacteria[138]. In addition, human cells can switch between a state of high cell proliferation with a damped circadian oscillator, to a state of low proliferation but robust circadian rhythms, depending on molecular interactions and activities of cell cycle and clock regulators[139].

Here, we exploit the fact that the two coupled cycles evolve on a low dimensional and compact manifold (the flat torus) to fully characterize their dynamics. In particular, starting from a generic stochastic model for the interacting phases combined with fluorescence microscopy recordings from thousands of individual cells, we obtained a data-driven reconstruction of the coupling function describing how the cell cycle influences the circadian oscillator. This coupling phase-locks the two oscillators in a temperature-independent manner, and only few of the deterministically predicted phase-locked states were stable against inherent fluctuations. Moreover, we established that the coupling between the two oscillators is conserved from mouse to human, and can override systemic synchronization signals such as temperature cycles. Finally, we showed in a physiological context how this coupling explains why mammalian tissues with different cell proliferation rates have shifted circadian phases.

2.3. Results

2.3.1. Modeling the dynamics of two coupled biological oscillators

To study the phase dynamics of the circadian and cell cycle oscillators, we reconstructed a stochastic dynamic model of the two coupled oscillators from single-cell time-lapse microscopy traces of a fluorescent Rev-erbα-YFP circadian reporter[45], [100]. Our approach consists in explicitly modeling the measured fluorescent signals, using a set of stochastic ordinary differential equations (SODEs) whose parameters are estimated by maximizing the probability of observing the data over the entire set of cell traces (Methods). We here present the key components of the model (detailed in Supplementary Information).

2.3.1.1. Phase model

First, we represent the phase dynamics of the circadian oscillator ($\theta = 0$ corresponds to peaks of fluorescence) and cell cycle ($\phi = 0$ is the cytokinesis) on a $[0,2\pi) \times [0,2\pi)$ torus. Since we showed previously that the influence of the clock on the cell cycle was negligible in

NIH3T3 cells[45], we here model only how the cell-cycle progression influences the instantaneous circadian phase velocity ω_θ using a general coupling function $F(\theta,\phi)$ (Figure 1.1a). To account for circadian phase fluctuations and variability in circadian period length known to be present in single cells[100], [140], we added a phase diffusion term $\sigma_\theta dW_t$. For the cell-cycle phase, we assumed a piecewise linear and deterministic phase progression in between two successive divisions. The SODEs for the phases read:

$$\begin{cases} d\theta &= \dfrac{2\pi}{T_\theta}dt + F(\theta,\phi)dt + \sigma_\theta dW_t \\ d\phi &= \dfrac{2\pi}{T_\phi^i}dt \end{cases} \quad (1)$$

Here, T_θ represents the intrinsic circadian period, while the term T_ϕ^i represents the i^{th} cell-cycle interval between two successive divisions.

2.3.1.2. Model of the signal

We linked the circadian phase with the measured time traces through a 2π-periodic function $w(\theta)$. In addition, as suggested by typical data traces (Supplementary Figure 1.1a), we considered amplitude (A_t) and baseline (B_t) fluctuations, which for simplicity we modeled as independent from θ_t, an assumption that was supported *a posteriori* (Supplementary Information). The full model for the observed signal S_t thus reads:

$$S_t = e^{A_t} w(\theta_t) + B_t + \xi \quad (2)$$

where ξ is a normally distributed random variable (measurement noise) and A_t, B_t are Ornstein-Uhlenbeck processes varying more slowly than the phase distortion caused by $F(\theta,\phi)$, *i.e.*, on timescales on the order of the circadian period (Supplementary Information).

2.3.1.3. Inference of phases & coupling function

From this stochastic model (Equations 1&2, Figure 1.1b), we built a Hidden Markov Model (HMM) to calculate posterior probabilities of the oscillator phases at each measured time point, using the forward-backward algorithm[141]. To estimate $F(\theta,\phi)$, we used a maximum-likelihood approach that combines goodness of fit with sparseness and smoothness constraints, which we implemented with an Expectation-Maximization (EM) algorithm (Methods, Supplementary Information).

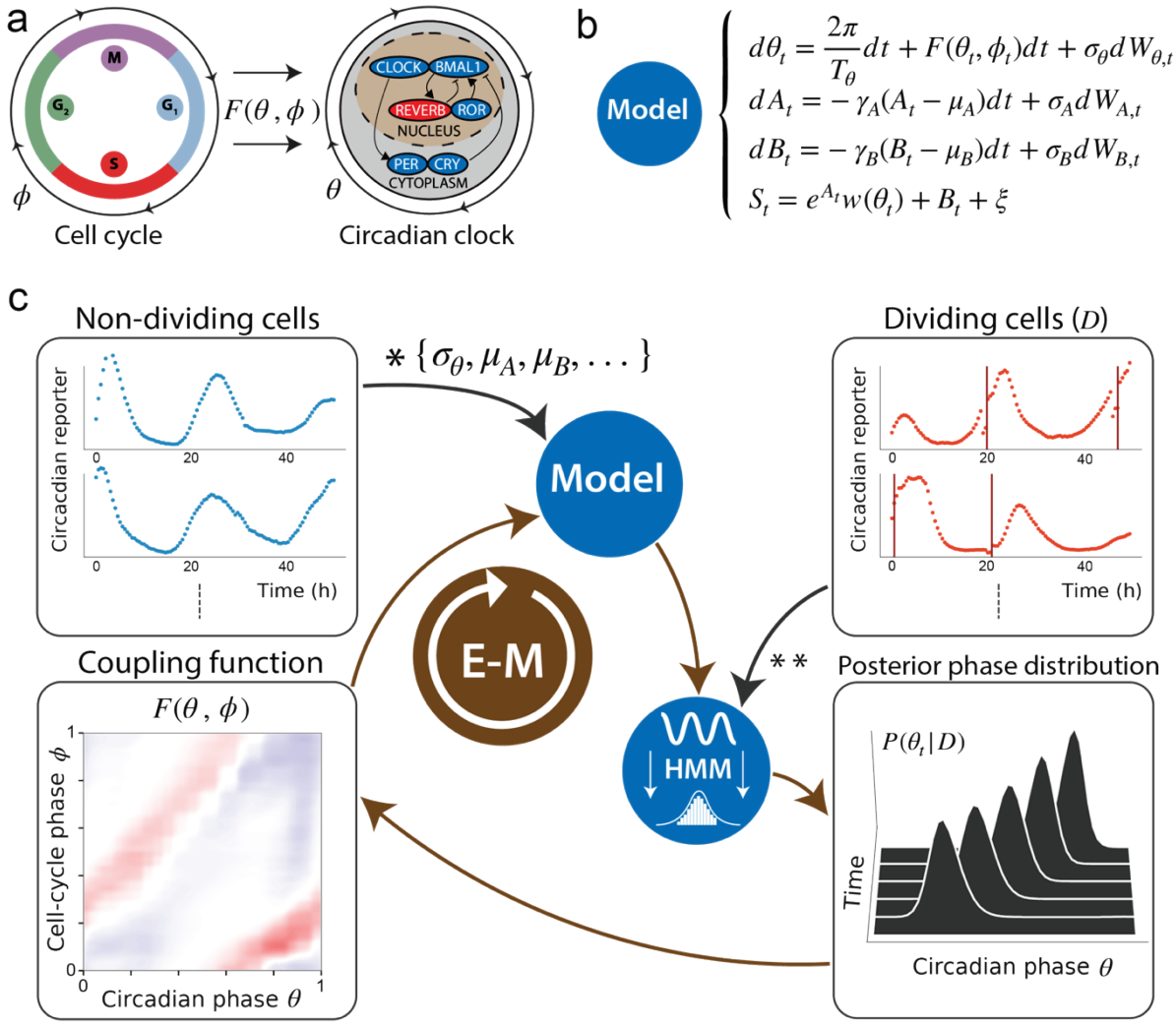


Figure 1.1: Reconstructing the phase dynamics and coupling of two biological oscillators. (**a**) In mouse fibroblasts, the cell cycle (left) can influence the circadian oscillator (right) according to a coupling function $F(\theta, \phi)$, where ϕ denotes the cell cycle and θ the circadian oscillator phases. (**b**) Stochastic model for the signal S_t using diffusion-drift SODEs for the circadian phase θ_t, amplitude A_t and background B_t fluctuations, as well as a function $w(\theta)$ linking the phase θ_t to the measured observations, and $F(\theta, \phi)$ (**c**) Fluorescence microscopy traces (Rev-erbα-YFP circadian reporter) are recorded for non-dividing and dividing cells (top left and top right boxes). Coupling-independent parameters (*) are estimated from non-dividing cells while dividing cells are necessary to infer $F(\theta, \phi)$(**). The optimization problem is solved by converting the model to a HMM in which θ_t, A_t and B_t are latent variables. The HMM is used on traces to compute posterior probabilities of circadian phases (bottom right box), while the cell-cycle phase is retrieved using linear interpolation between successive divisions (top right box, vertical orange lines). An iterative EM algorithm then yields the converged $F(\theta, \phi)$ (bottom left box).

The successive steps of our approach are illustrated in Figure 1.1c. Dividing cells indicated that, typically, the circadian phase progression shows variations in phase velocity (Supplementary Figure 1.1a). To validate that these variations can identify $F(\theta, \phi)$, we generated noisy traces *in silico* with pre-defined $F(\theta, \phi)$ and reconstructed the coupling function, showing excellent qualitative agreement (Supplementary Figure 1.1b-c).

2.3.2. Influence of the cell cycle on the circadian phase

In mouse embryonic fibroblasts (NIH3T3), we showed that due to the coupling, circadian periods decrease with temperature in dividing cells, but not in quiescent cells[45]. To further understand how temperature influences the interaction between the two oscillators, we re-analyzed NIH3T3 traces (24-72h long) obtained at 34°C, 37°C, and 40°C[45]. From those, we found that both the inferred coupling functions and phase densities at the three temperatures were very similar, with almost identical 1:1 phase-locked orbits (Supplementary Figure 1.2a-c). We therefore modeled the coupling as temperature-independent and re-constructed a definitive $F(\theta, \phi)$ from traces at all temperatures (Figure 1.2a, Supplementary Figure 1.2d). This function shows a diffuse structure mainly composed of two juxtaposed diagonal stripes: one for phase acceleration (red), and one, less structured, for deceleration (blue). The slopes of these stripes are about one, which indicates that an approximate minimal model of the coupling would be a function $F(\theta, \phi) = f(\theta - \phi)$. However, the phase velocity varies along the stripes and attractor (see below), which justifies using a 2D parameterization of the coupling function. The phase density for cells with fixed cell-cycle period of 22h (corresponding to the mean cell-cycle period in the full dataset) (Figure 1.2b, and Movie 1.1) clearly suggests 1:1 phase-locking. In fact, analyzing the predicted deterministic dynamics (Equation 1, with the reconstructed $F(\theta, \phi)$, and without the noise) shows a 1:1 attractor (Figure 1.2c). Thus, in this 1:1 state, the endogenous circadian period of 24h is shortened by two hours, which results from acceleration occurring after cytokinesis ($\phi = 0$) when the circadian phase is near $\theta \approx 0.8 \times 2\pi$, and lasting for the entire G1 phase, until about $\theta \approx 0.4 \times 2\pi$ when cells typically enter S phase ($\phi \approx 0.4 \times 2\pi$ at the G1/S transition, Supplementary information, Figure 1.2d).

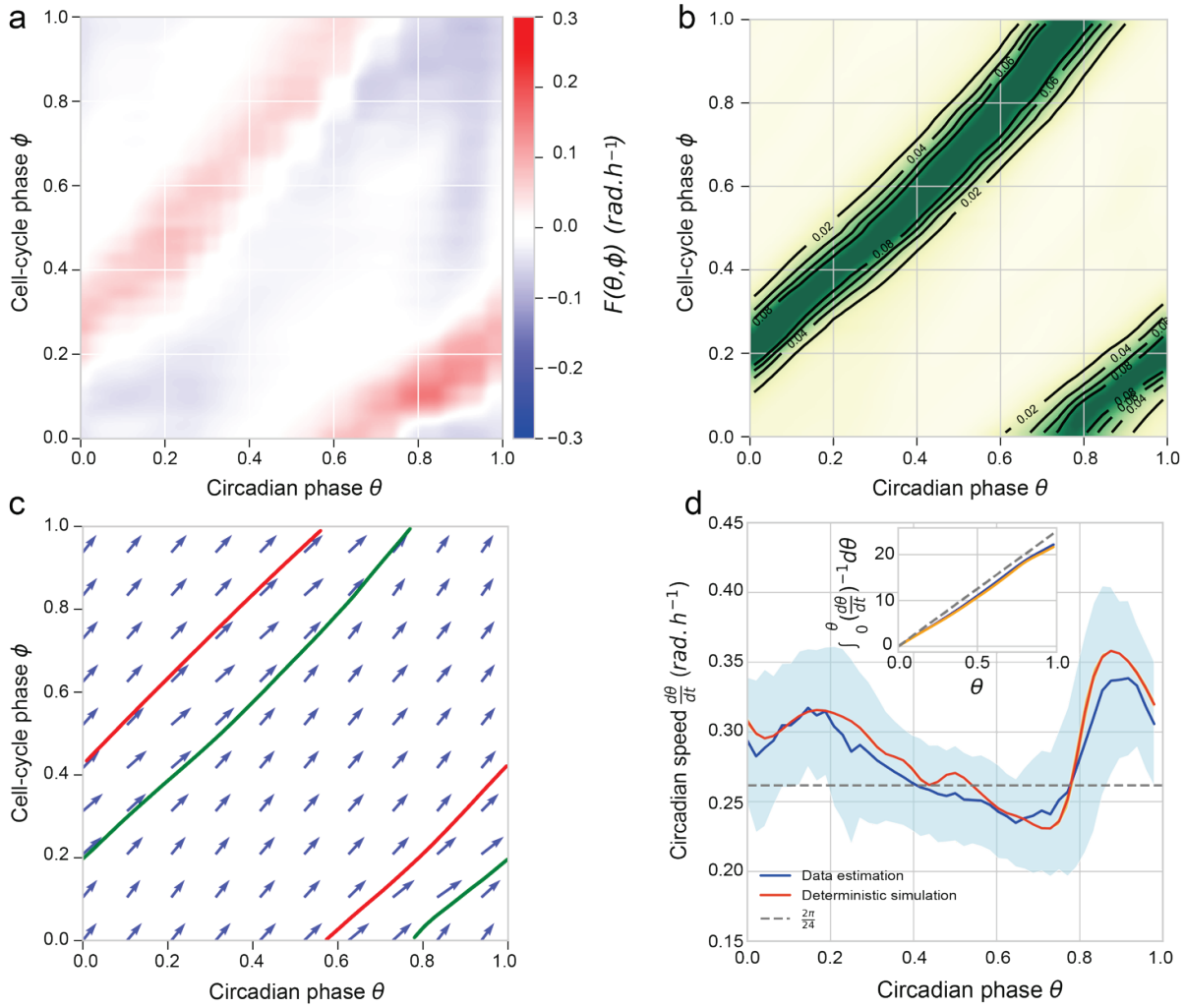


Figure 1.2: Influence of the cell cycle on the circadian phase enables 1:1 phase-locking. (**a**) Coupling $F(\theta,\phi)$ optimized on dividing single-cell traces. Due to similar results (Supplementary Figure 1.2) traces from the three temperatures (n=154, 271, 302 traces at 34°C, 37°C, 40°C, resp.) are pooled. (**b**) Density of inferred phase traces from all the dividing traces with 22±1h cell-cycle intervals indicates a 1:1 phase-locked state. (**c**) Numerical integration of phase velocity field (arrows, deterministic model) yields 1:1 attractor (green line) and repeller (red line). Here, the cell-cycle period was set to 22 h. (**d**) Circadian phase velocity is not constant along the attractor, here for cells with 22±1h cell-cycle intervals. Data (blue line, standard deviation in light-blue shading) and deterministic simulation (orange line). Inset: integrated time along the attractor. The gray line shows constant bare phase velocity $\omega_\theta = \frac{2\pi}{24h}$.

2.3.3. Phase dynamics in perturbation experiments

The reconstructed model allows us to simulate the circadian phase dynamics in function of the cell-cycle period, which is relevant as the cells display a significant range of cell-cycle lengths (Supplementary Figure 1.3a). In the deterministic system, we find 1:1 phase-locking over a range of cell-cycle times varying from 19h to 27h, showing that the cell cycle can both globally accelerate and slow down circadian phase progression (**Figure 1.3**a). The attractor shifts progressively to the right in the phase-space, yielding a circadian phase at division ranging from $\theta \approx 0.7 \times 2\pi$ at division when $T_\phi = 19h$ to $\theta \approx 0.9 \times 2\pi$ when $T_\phi =$

$27h$. Since the attractor for different cell-cycle periods shifts, the circadian phase velocity profile also changes (Supplementary Figure 1.3b). To validate the predicted shifts, we experimentally subjected cells to perturbations inducing a large variety of cell-cycle periods and compared the observed circadian phase to the model prediction at three different cell-cycle phases, revealing an excellent agreement, with no additional free parameters (**Figure 1.3**b).

The simulations also clearly revealed multiple phase-locked states (1:2, 1:1, 2:1, 3:1, etc, p:q indicating the number of cell cycles p and the number of clock cycles q), represented as Arnold tongues (**Figure 1.3**c, and Movie 1.2 for an animated phase-space representation). We identified cell data trajectories following almost perfectly the deterministic attractors, both in the 1:1 and 1:2 phase-locking states (**Figure 1.3**e and f, respectively); however, cells showing other p:q states were rarely observed.

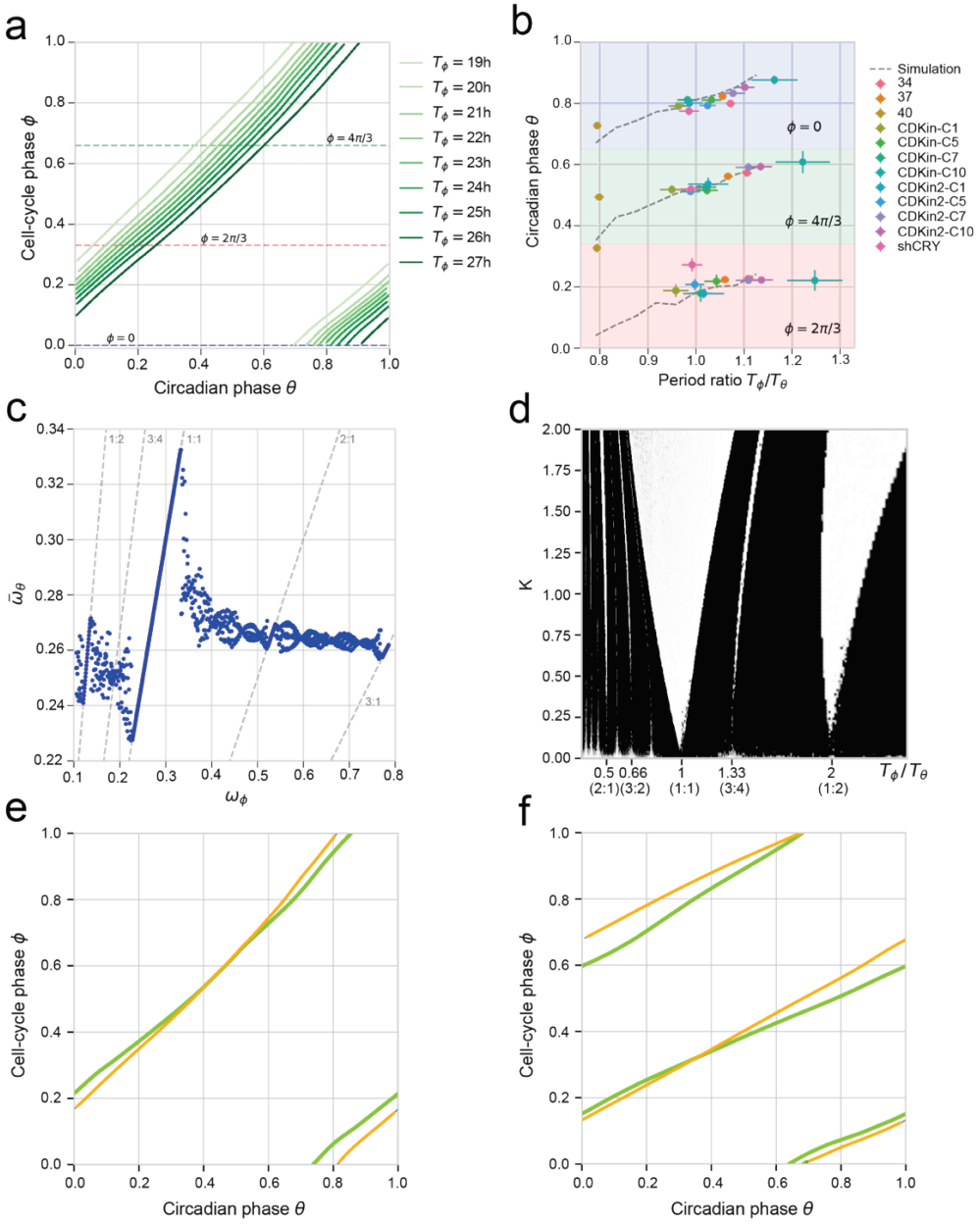


Figure 1.3: The coupling between the cell cycle and the circadian oscillator predicts phase shifts and phase-locking attractors in perturbation experiments. (**a**) Simulated (deterministic) attractors for cell-cycle periods ranging from 19h to 27h show that the dephasing of the cell cycle and the circadian oscillator changes within the 1:1 state. Periods just outside of this range yield quasiperiodic orbits. The horizontal dashed lines indicate three different cell-cycle phases $\phi = 0, \phi = \frac{1}{3} \times 2\pi, \phi = \frac{2}{3} \times 2\pi$ used in panel (**b**). (b) Predictions from a) (dashed grey lines) against independent experimental data collected from 12 perturbation experiments (colored symbols, see legend, notation explained in Methods). (**c**) Multiple phased-locked states are predicted, recognizable by rational relationships between the frequencies of the entraining cell cycle and the entrained circadian oscillator, interspersed by quasiperiodic intervals. (**d**) Arnold tongues showing multiple phase-locked states in function of cell-cycle periods and coupling strength (K=1 corresponds to the experimentally found coupling). Stable zones (white tongues) reveal attractors interspersed by quasi-periodicity. Although there are only two wider phased-locked state (1:1 and 1:2), several other p:q states are found. (**e-f**) Representative single-cell traces (data in yellow) evolving near predicted attractors (green lines). A cell with $T_\phi = 24h$ (e) and one with $T_\phi = 48h$ (f) near the 1:1 and 1:2 orbits, respectively.

2.3.4. Fluctuations extend 1:1 phase-locking asymmetrically

To understand the differences between the simulated deterministic system and observed cell traces, we simulated the stochastic dynamics (Equation 1). We then compared measured data trajectories stratified by cell-cycle periods (**Figure 1.4**a) with deterministic (**Figure 1.4**b) and stochastic simulations (**Figure 1.4**c). This revealed that data agree better with stochastic than deterministic simulations, indicating that the phase fluctuations qualitatively change the phase portrait. One striking observation is the increased range of 1:1 phase-locking in the noisy system, however asymmetrically, since this occurs for shorter, but not for longer cell-cycle periods. Indeed, while 1:2 phase-locking is observed in the data and the noisy simulations, the deterministically predicted 2:1 state is replaced in the data and stochastic system by 1:1-like orbits. Consistently, spectral analysis revealed significant differences between deterministic and stochastic simulations (**Supplementary Figure 1.4**a-b, Movie 1.3); in addition, the coupling, specifically in the 1:1 state, was able to efficiently filter the noise (**Supplementary Figure 1.4**b, right).

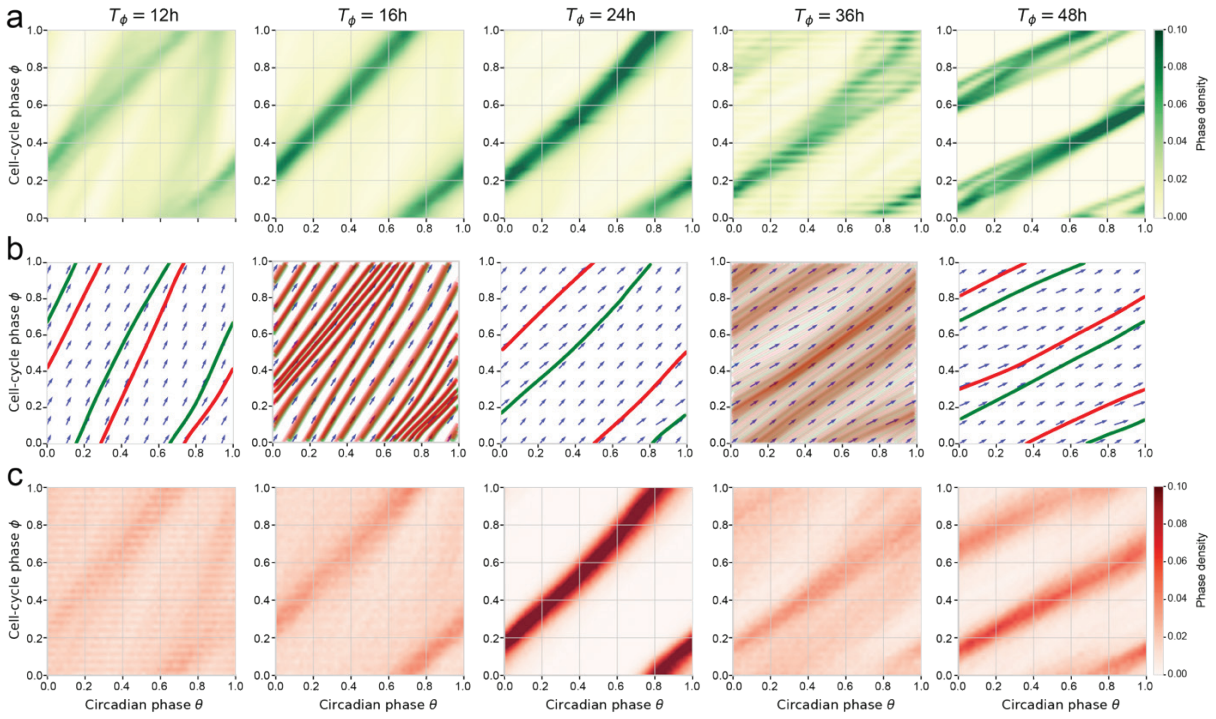


Figure 1.4: Single-cell data and stochastic simulations reveal robust 1:1 and 1:2 phase-locked states. (**a**) Phase-space densities from the experimental traces stratified by cell-cycle periods (±1h for each reference period); n=16, 223, 303, 54, 4 cell traces in the T=12, 16, 24, 36, 48 h panels, resp. (**b**) Vector fields and simulated (deterministic) trajectories for the different cell-cycle periods. Attractors are shown in green (forward time integration) and repellers (backward integration) in red (see also Movie 1.2). (**c**) Phase-space densities obtained from stochastic simulations of the model match better with the data compared to b.

2.3.5. Evolutionarily conserved phase-locking

Most studies investigating the interaction between the cell cycle and the circadian oscillators in mammals are from rodent models[44], [45], [100], [110], [111], [119]. To test if the above phase-locked dynamics are conserved in human U2OS cells, an established circadian oscillator model[142], [143], we engineered a U2OS cell line termed U2OS-Dual expressing a dual circadian fluorescent (Rev-erbα-YFP) and luminescent (Bmal1-luc) reporter system. U2OS-Dual cells possess a functional circadian clock behaving similarly to NIH3T3 cells also expressing a Bmal1-Luc luminescent reporter[144] (**Figure 1.5**a). We scrutinize the relation between the two cell lines by comparing their behavior in different conditions: at 34°C and 37°C for cells with synchronized and non-synchronized circadian cycles[45] (**Figure 1.5**b-g).

Similarly to NIH3T3 cells, the division events of non-synchronized U2OS-Dual cells grown at 37°C occurred 4.96 ± 2.6 h before a peak in the circadian fluorescent signal (**Figure 1.5**c), indicating that the cell cycle and the circadian clock interact. To investigate the directionality of this interaction, we tested, like in NIH3T3 cells[45], whether the circadian clock phase could influence cell-cycle progression by resetting the circadian oscillator using dexamethasone (dex), a circadian resetting cue[145] that does not perturb the cell cycle[100]. We found the expected resetting effect of dex on the circadian phase by the density of peaks in reporter levels during the first 10h of recording, but with unnoticeable effects on the timing of the first division (**Figure 1.5**d). However, the circadian peak following the first division occurred systematically around 5h after the division in both conditions, suggesting that cell division in U2OS can reset circadian phases and overwrite dex synchronization. Synchronization of the circadian clocks for dex- *vs* non- treated cells was expectedly higher for dex and gradually decreased to reach the level of the untreated cells (**Figure 1.5**e), contrasting with the generally lower synchronization of cell divisions in both conditions. To then test if the cell cycle could influence the circadian clock, we lengthened the cell-cycle period by growing cells at 34°C and compared with 37°C. Interestingly, cells at 34°C showed a longer circadian period compared to 37°C (**Figure 1.5**f), unlike the temperature compensated circadian periods (~25h) in non-dividing cells (**Figure 1.5**a, g-h).

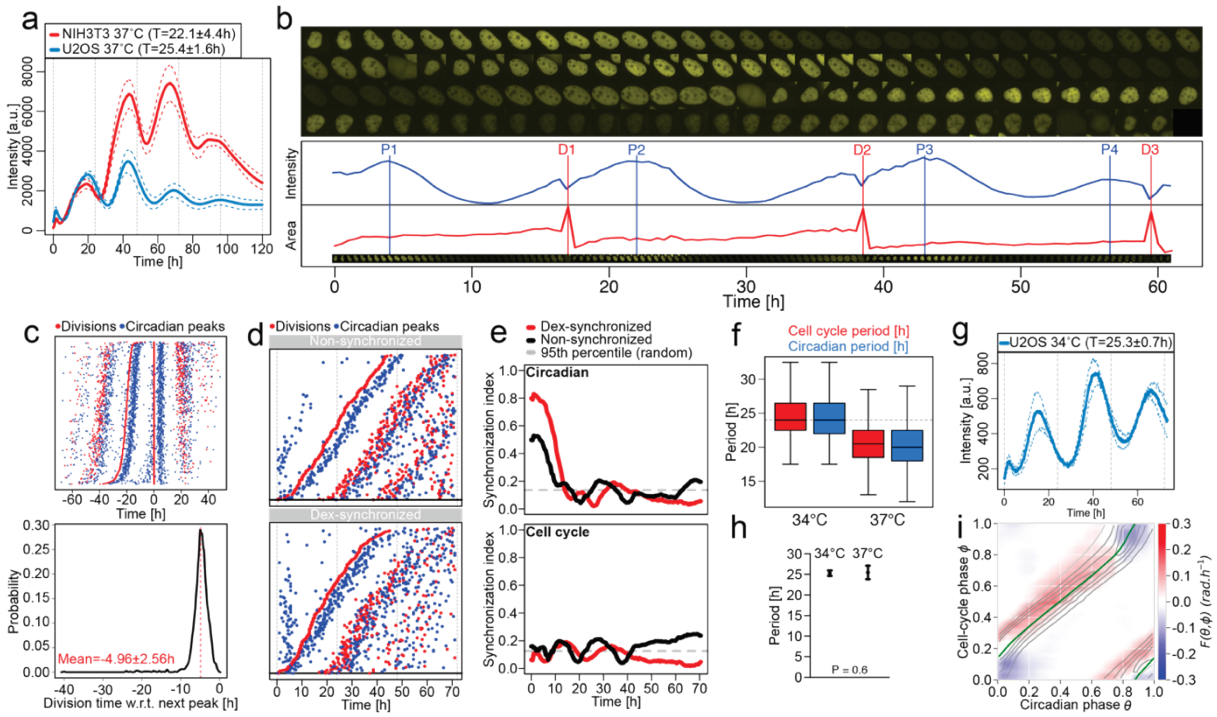


Figure 1.5: Conserved influence of the cell cycle on the circadian clock in human U2OS osteosarcoma cell. (**a**) Mean luminescence intensities (±SD, n=3) from non-dividing NIH3T3 and U2OS cells grown at 37°C expressing a Bmal1-Luc reporter. Values in the legend correspond to the mean periods ±SD. (**b**) Semi-automated segmentation and tracking of U2OS cell lines expressing the Rev-erbα-YFP circadian fluorescent reporter. Red vertical lines represent cell divisions (cytokinesis) and blue vertical lines show Rev-erbα-YFP signal peaks. (**c**) Top: stack of divisions (red) and Rev-erbα-YFP peaks (blue) for single U2OS traces centered on divisions. Bottom: distribution of the time of division relative to the next Rev-erbα-YFP peaks (in read mean±SD, n=1298). (**d**) Divisions and Rev-erbα-YFP peaks from single non-synchronized (top), and dexamethasone (dex)-synchronized (bottom) U2OS traces ordered on the first division. (**e**) Synchronization index (SI) from non-synchronized (black) and dex-synchronized (red) traces for the circadian phase (top) and cell-cycle phase (bottom) estimated as in Bieler et al.[45]. The circadian SI from non-synchronized cells is relatively high due to plating. Dashed gray lines show 95$^{\text{th}}$ percentiles of the SI for randomly shuffled traces. (**f**) Cell-cycle and circadian periods for U2OS cells grown at 34°C and 37°C (n > 90 for all distributions). (**g**) Mean luminescence intensities (±SD, n=3) for non-dividing U2OS cells grown at 34°C expressing a Bmal1-Luc reporter. Values in the legend correspond to the mean periods ±SD. (**h**) Mean and standard deviation of the circadian period for non-dividing U2OS cells grown at 34°C and 37°C (n=8 at 34 and n=9 at 37, two-sided Wilcoxon's test). (**i**) Coupling function $F(\theta,\phi)$ optimized on n=551 dividing U2OS traces grown at 37°C, superimposed with the attractor (T_ϕ=22h) obtained from deterministic simulations (green line).

Thus, similarly to mouse NIH3T3 cells, the coupling directionality is predominantly from the cell cycle to the circadian clock. In fact, the reconstructed coupling function for U2OS-Dual cells grown at 37°C (**Figure 1.5**i) is structurally similar to that obtained in mouse fibroblasts (Figure 1.2a), with the ensuing dynamics also showing a 1:1 attractor.

2.3.6. Dividing cells lose circadian temperature entrainment

In mammals, circadian clocks in tissues are synchronized by multiple systemic signals[99]. In fact, temperature oscillations mimicking those physiologically observed can phase-lock circadian oscillators in non-dividing (contact-inhibited) NIH3T3 cells *in vitro*[146]. To study how the found interactions influence temperature entrainment, we applied temperature cycles (24h period ranging from 35.5°C to 38.5°C) to U2OS cells growing at different rates (plated at different densities) and monitored population-wide Bmal1-luc signals (**Figure 1.6**a). We found that, independently of initial densities, as the populations reach confluency, the phases and amplitudes become stationary, showing 1:1 entrainment (**Figure 1.6**b, c and Supplementary Figure 1.5a). During the initial transients, emerging circadian oscillations in non-confluent cells showed phases that were already stationary, at least once cell numbers were sufficiently high to obtain reliable signals.

As cell confluence increases, the proportions of cells which stop cycling (exit to G0) increases[147]. We therefore hypothesized that the observed phase and amplitude profiles in Bmal1-luc signals (**Figure 1.6**b, c) originate from a mixture of two populations: an increasing population of non-dividing cells (G0) showing 'normal' entrainment properties, and dividing cells. We considered three scenarios for the dividing cells: i) the circadian oscillators in dividing cells adopt the same circadian profile as non-dividing entrained cells; ii) are not entrained; or iii) are entrained, but with a different phase compared to non-dividing cells (Supplementary Figure 1.5b). These scenarios can be distinguished by the predicted phase and amplitude profiles (Supplementary Figure 1.5c). Clearly, the measured profiles for U2OS-Dual cells favored the second scenario, suggesting that circadian oscillators in dividing cells do not entrain to the applied temperature cycles.

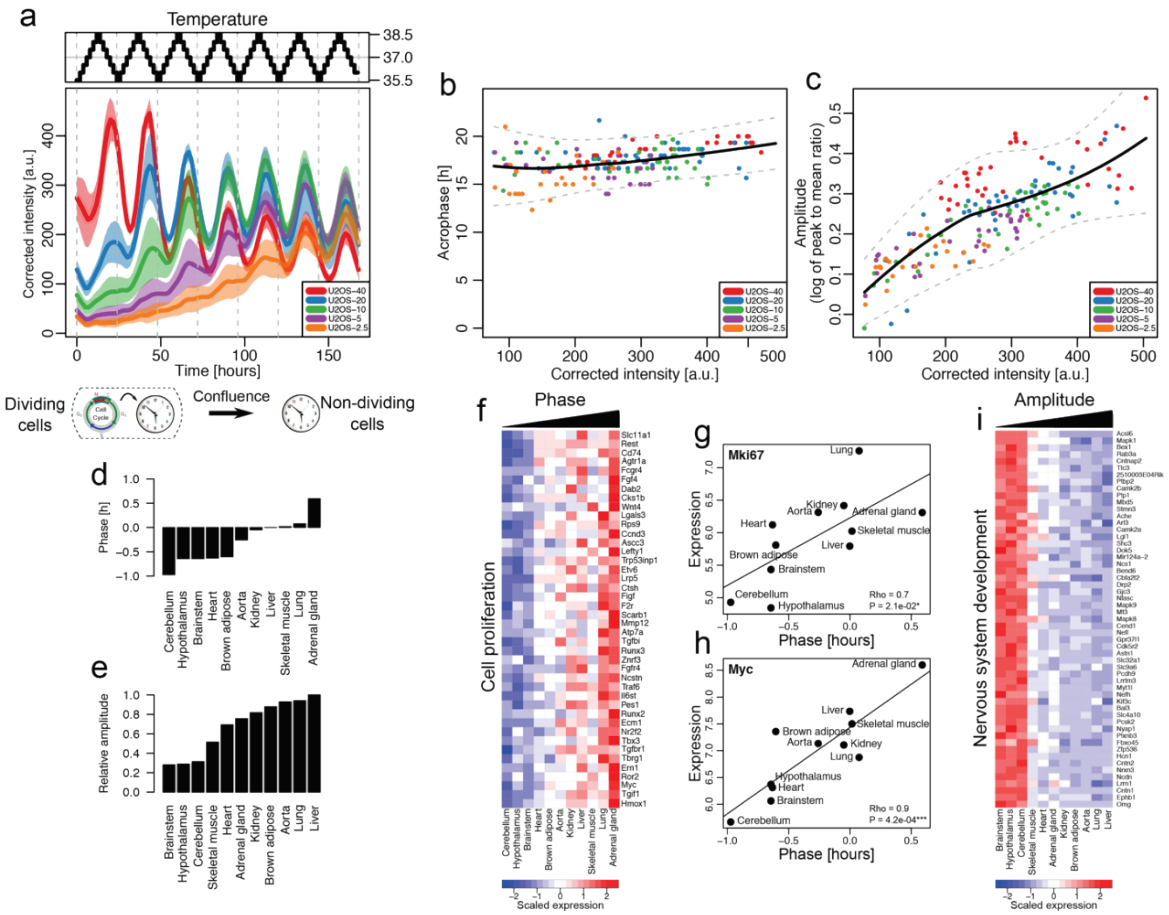


Figure 1.6: Temperature cycles do not entrain circadian oscillators in dividing cells and proliferation genes are associated with tissue-specific circadian phases. (**a**) Corrected and averaged Bmal1-Luc intensities and 95% confidence intervals (n=6) from U2OS-Dual cells plated at different initial densities and subjected to a temperature entrainment (top). (**b**) Acrophases (times of the local peaks in luminescence) of the Bmal1-Luc signal in function of the reporter intensity for the cells in a). Loess fit (black) and 95% confidence intervals (gray). (**c**) Amplitude (log of peak to mean ratio) of the Bmal1-Luc oscillations in function of the reporter intensity for the cells in a). Loess fit (black) and 95% confidence intervals (gray). (**d-e**) Circadian phases (d) and amplitudes (e) of different mouse tissues obtained in reference [148], relative to liver. (**f**) Expression levels of genes positively associated with phases from d) and linked to cell proliferation. (**g-h**) Correlations between Mki67(g) and Myc (h) mRNA expression and circadian phases across mouse tissues (Pearson's correlation, two-sided P-values from t-distribution with n-2 degrees of freedom). (**i**) Expression levels of genes negatively associated with amplitudes and linked to nervous system development.

2.3.7. Proliferation is associated with tissue-specific circadian phases

The above findings suggest that phases or amplitudes of circadian clocks in organs *in vivo* might be influenced by the proliferation state of cells in the tissue. To test this, we investigated circadian clock parameters in different mouse tissues using a study of mRNA levels in twelve adult (6 weeks old males) mouse tissues, which revealed that clock phases

span 1.5 hours between the earliest and latest tissues (**Figure 1.6**d)[148], [149], an effect which is considered large in chronobiology as even period phenotypes of core clock genes are often smaller[131], [150], [151]. We noticed that the mean mRNA levels across tissues of many genes correlated with the phase offsets (Supplementary Table 1.1). However, gene functions related to cell proliferation stood out as the most strongly enriched (**Figure 1.6**f, Supplementary Table 1.1). Among those genes, the levels of known markers of cell proliferation such as *Mki67* or *Myc* were strongly correlated with the phase offsets (**Figure 1.6**g-h, Supplementary Table 1.1). Amplitudes, on the other hand, were not correlated with proliferation genes, but rather with neuronal specific genes, as expected owing to the damped rhythms present in those tissues (**Figure 1.6**i, Supplementary Table 1.1)[149]. Thus, this analysis suggests that the differences in basal proliferation levels observed in normal tissues might underlie the dephasing of the circadian clock, suggesting a physiological role for the interaction of the cell cycle and circadian clocks.

2.4. Discussion

A goal in quantitative single-cell biology is to obtain data-driven and dynamical models of biological phenomena in low dimensions. In practice, the heterogeneity and complex physics underlying the emergence of biological function in non-equilibrium living systems, as well as the sparseness of available measurements pose challenges. Here, we studied a system of two coupled biological oscillators, sufficiently simple to allow data-driven model identification, yet complex enough to exhibit qualitatively distinct dynamics, *i.e. p:q* states and quasi-periodicity.

In the coupled cell cycle and circadian oscillator system, phase-locked states different from 1:1 have been observed[112]. While multiple attractors, notably 1:1 and 3:2, were found in mouse NIH3T3 cells under transient dexamethasone stimulation[44], we here report 1:2 states for long cell-cycle times under steady, unstimulated, conditions. Unlike other deterministically predicted states, 1:2 was sufficiently robust and observed in some cells. In fact, we found that noise extended the range of the 1:1 tongue, but asymmetrically, *i.e.* towards decreased cell-cycle periods. This may be reminiscent of generalized Poincaré oscillators showing that the entrainment range is broader for limit-cycles with low relaxation rates[152]. Indeed, noise could decrease relaxation rates and thereby broaden Arnold tongues. In addition, for certain cell cycle periods, we observed the superposition of multiple states, both in the data and in the stochastic simulations, which were not present in the deterministic analysis (**Figure 1.4**, see notably T=12h and T=36h). This is reminiscent of mode hopping as described in the context of an oscillatory gene circuit underlying inflammatory responses[71],

however, here the corresponding Arnold tongues do not overlap in the range of the biologically relevant coupling strength (K=1, **Figure 1.3**d).

While we focused on the emergent dynamics in the coupled oscillator system, considerations on possible biological mechanisms are relevant for follow-up biochemical analyses. How chromosome condensation or nuclear envelope breakdown may influence the circadian clock phase progression via either transcriptional shutdown, or displacement of chromatin bound circadian repressors, respectively, was discussed previously[45]. For example, *Rev-erbα* transcription being so tightly locked to cell divisions (the peak accumulation of the reporter occurs 5h after mitosis) could reflect the sudden derepression of its promoter, due to displaced *CRYPTOCHROME1* (CRY1) containing repressor complex following nuclear envelope breakdown[153]. In turn, REV-ERB-A accumulation influences clock phase by binding to promoters of multiple core clock components, including *Cry1*[154], [155]. More specific transcriptional activities could also play a key role in coupling the cell and circadian cycles. In fact, the circadian oscillator is exquisitely sensitive to numerous signaling pathways, impinging upon the clock by transcriptional induction of *Period* genes, which thus provides an efficient synchronization method[145]. Similarly, entrainment via temperature cycles also converges onto *Period* gene transcription[100]. However, we are not aware of cell-cycle dependent transcriptional regulators, such as E2F factors, targeting clock components like the *Period* genes. Finally, since the regulation of protein stability is important for clock function[156], it is possible that phosphorylation-controlled proteolytic activities driving the cell cycle could target circadian phase regulators[157], thereby mediating the observed coupling.

In mammals, the circadian oscillator in the suprachiasmatic nucleus (SCN) is the pacemaker for the entire organism[158], driving 24h rhythms in activity, feeding, body temperature and hormone levels. In particular, the SCN can synchronize peripheral cell-autonomous circadian clocks located within organs across the body[159]. Consistent with theory[70], in a physiological context of entrainment, the coupling of the cell cycle with the circadian clock can induce proliferation dependent phase-shifts, which we observed. Such phase shifts could reflect a homogenous behavior of all cells, or it could reflect heterogeneity of cell proliferation states, possibly leading to wave propagation. The phase shifts we observed in tissues were associated with low proliferation, *i.e.* non-pathological states of tissue homeostasis and cell renewal. For example, the liver or the adrenal gland showed a phase advance compared to fully quiescent tissues like the brain.

When cell proliferation is abnormally high such as in cancer, circadian clocks are often severely damped[160]. While this absence of a robust circadian rhythm in malignant tissue states may reflect non functional circadian oscillators due to mutations in clock genes[161], the damped rhythms may also reflect circadian desynchrony of otherwise functional circadian

oscillators. Such desynchrony would readily follow from the coupling between the cell-cycle and circadian oscillators we highlight here, in the presence of non-coherent cell-cycle progression.

Methodologically, the new approach to reconstruct a dynamical model for the coupled oscillator system has significant advantages over previous methods, notably strong assumptions such as the sparse and localized coupling are dispensable[45]. Compared with generic model identification techniques[162], our approach models the raw data and its noise structure explicitly. In the future, such data-driven identification of dynamical models might reveal dynamical instabilities underlying ordered states in spatially extended systems, as occurring, for instance, during somitogenesis[163].

2.5. Methods

2.5.1. Cell lines

All cell lines (U2OS-Dual, NIH3T3-Bmal1-Luc, and U2OS-PGK-Luc) were maintained in a humidified incubator at 37°C with 5% CO2 using DMEM cell culture media supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin-glutamine (PSG). One day before luminescence or fluorescence acquisitions, we replaced DMEM with FluoroBrite DMEM media supplemented with 10% FBS and 1% PSG. NIH3T3 perturbation experiments were generated in Bieler et al[45]. Briefly, they correspond to temperature changes (34, 37, and 40 ˚ C), treatment with CDK1 (RO-3306, Sigma-Aldrich) and CDK1/2 (NU-6102, Calbiochem) inhibitors at 1, 5, 7 and 10µM (CDK1in-[1,5,7,10] and CDK2in-[1,5,7,10]), and shRNA-mediated knockdown of Cry2 (shCry).

2.5.2. Fluorescent time-lapse microscopy

Time-lapse fluorescent microscopy for U2OS-Dual cells was performed at the biomolecular screening facility (BSF, EPFL) using an InCell Analyzer 2200 (GE healthcare). Experiments were performed at different temperature (34°C, 37°C, or 40°C) with a humidity and CO_2 (5%) control system. We used 100 ms excitation at 513/17 nm and emission at 548/22 nm to record the YFP channel. Cells were recorded by acquiring one field of view per well in a 96-well black plate (GE healthcare). We used our previously developed semi-automated pipeline for segmentation and tracking of individual cells[45]. In total, traces from n=551 U2OS cells were obtained (typically 50 cells are obtained per movie). NIH3T3 single cells trace are reanalyzed from previous work[45]; here, we used n=2504 of those time traces. In all cases (NIH3T3 and U2OS), we followed several quality control metrics from [45]. Briefly, we discarded all traces that left the field-of-view at some point during the acquisition.

We also visually inspected all traces, using a custom-made Matlab tool, to remove traces with problematic segmentation and tracking. In addition, we only kept traces with significant circadian amplitude (peak height >0.25, rescaled signals, Supplementary Information). To minimize boundary artifacts, typically, only traces with at least 2 full cell cycles were kept. The number of cells used for specific analyses, including sub-selections of traces based on the cell-cycle intervals, are indicate in the figure captions.

2.5.3. Inferring the phase dynamics of two biological oscillators

Denote by $\boldsymbol{D}$ the entire set of single cells traces comprising temporal intensity measurements (Δt=30 min) from all fluorescent traces and $\boldsymbol{\Lambda}$ the set of model parameters, comprising the gridded coupling function F_{ij}. Note that all parameters are shared by all cells in $\boldsymbol{D}$. To reconstruct the phase dynamics of our model, we seek to maximize the likelihood of the data $\mathcal{L}(\boldsymbol{\Lambda}|\mathbf{D})$, that is, we solve:

$$\boldsymbol{\Lambda}^* = \operatorname{argmax} \mathcal{L}(\boldsymbol{\Lambda}|\mathbf{D}) \qquad (i)$$

In practice, we used an EM algorithm, by iteratively optimizing the function $Q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}')$ over its first argument, where Q can be written as follow:

$$Q\left(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{'}\right) = E\left[\log p(\mathbf{D}, \mathbf{X}|\boldsymbol{\Lambda})\middle|\mathbf{X}, \boldsymbol{\Lambda}^{'}\right] \qquad (ii)$$

That is, $Q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}')$ corresponds to the expected value of the log-likelihood of the data with respect to the posterior probabilities of the hidden phases $\mathbf{X}$ (latent variables), computed using the current parameter $\boldsymbol{\Lambda}'$. This process guarantees a monotonous convergence of the log-likelihood, although a global maximum is not necessarily reached[44].

To control for the many parameters F_{ij}, we added regularization constraints for both the smoothness and sparsity:

$$Q_p(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') = Q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') - \lambda_1 \sum_{ij} ||F_{ij}||^2 - \lambda_2 \sum_{ij} F_{ij}^2 \qquad (iii)$$

This expression is also guaranteed to converge[128].

Details about the optimization method, choice of the regularization parameters, and computation of the phase posteriors using a HMM are provided in Supplementary Information.

2.5.4. Long-term temperature entrainment and luminescence recording

We performed long-term temperature entrainment experiments using a Tecan plate reader Infinite F200 pro with a CO_2 and temperature modules. One day before starting the experiment, serial dilution ranging from 40,000 to 2,500 cells were seeded in 96-well white flat bottom plates (Costar 3917). To prevent media evaporation, all wells were filled with 300µl of media composed of FluoroBrite, 10% FBS, 1% PSG, and 100 nM D-luciferin (NanoLight technology) and covered with a sealing tape (Costar 6524). We set up temperature entrainment using stepwise increase (or decrease) of 0.5°C every 2 hours to produce temperature oscillating profiles going from 35.5°C to 38.5°C and back to 35.5°C again over a period of 24 hours. Intensities from all wells were recorded every 10 minutes with an integration time of 5000 milliseconds. Since temperature impacts the enzymatic activity of the luciferase[164], we corrected the signal for this systematic effect (Supplementary Information).

2.5.5. Association between gene expression and phase in tissues

We used the average gene expression obtained from a selected set of twelve adult (6 weeks old males) mouse tissues from the Zhang et al. dataset (GEO accession GSE54650)[149]. For this analysis, we estimated the Pearson's correlation between the averaged gene expression and the circadian tissue phases or amplitudes reported in reference [148]. We selected the top 200 genes positively or negatively associated with either the phases or the amplitudes for gene ontology analysis[165] (Supplementary Table 1.1).

2.6. Data availability

The data supporting figures and other findings of this study are available from the corresponding author on request.

2.7. Code availability

The code is available online at the following URL:

https://github.com/ColasDroin/CouplingHMM

2.8. Acknowledgments

We thank Rosamaria Cannavo for engineering the U2OS-Dual cell line and Jonathan Bieler for initial analyses. Fabien Kuttler from the Swiss Federal Institute of Technology

(EPFL) Biomolecular Screening Facility (EPFL-BSF) and Luigi Bozzo and José Artacho from the EPFL Bioimaging and Optics Core Facility (EPFL-BIOP) for assistance with the imaging. Fluorescence-activated cell sorting was performed at the EPFL Flow Cytometry Core Facility (EPFL-FCCF). This work was supported by the Swiss National Science Foundation Grant 310030_173079 and the EPFL. E.R.P was supported by a Canadian Institute of Health Research (CIHR 358808) and a SystemsX.ch Transition Postdoc Fellowships (51FSP0163584).

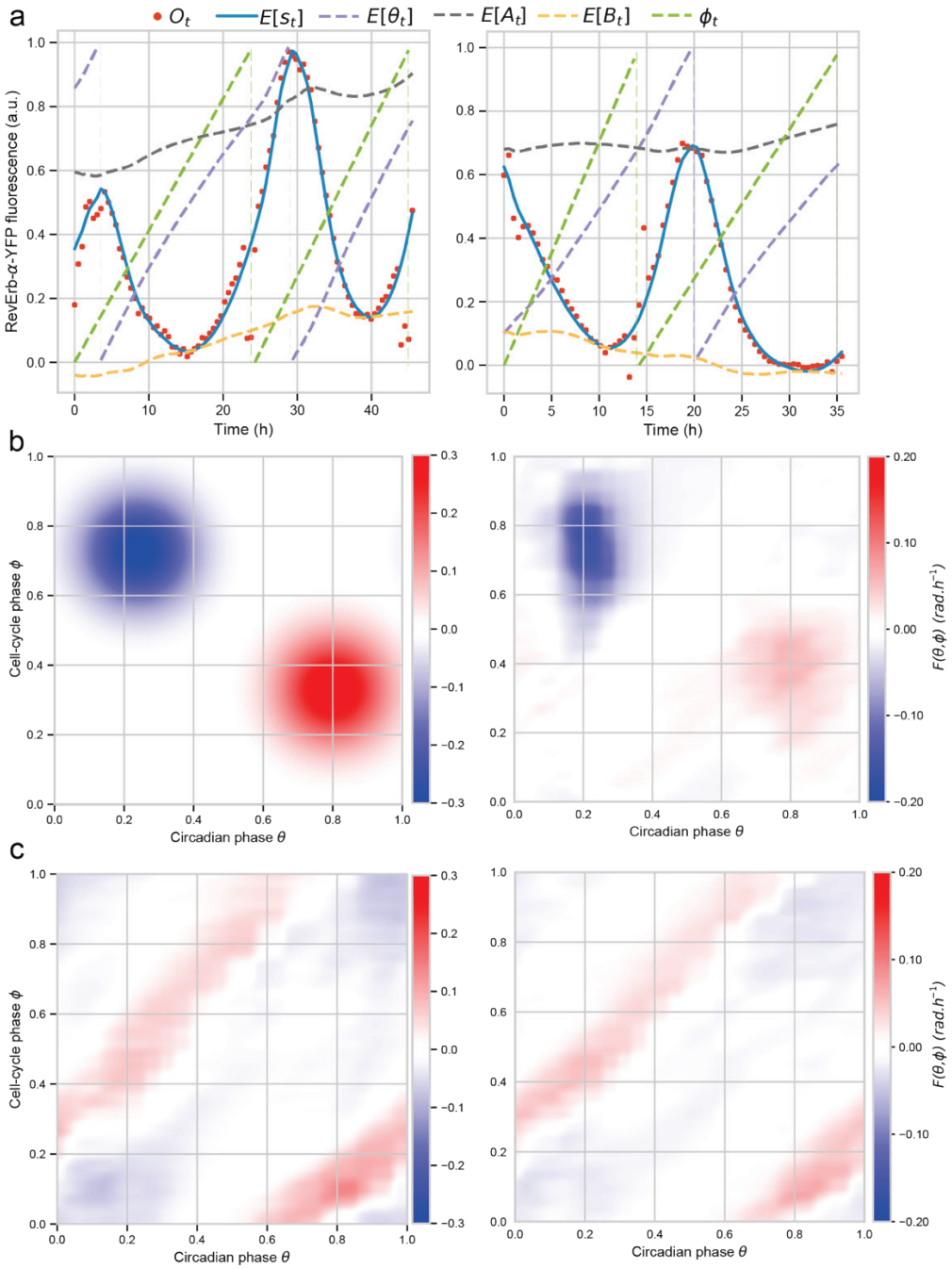
2.9. Contributions

C.D., E.R.P., and F.N. designed and participated in the study concept. C.D. and E.R.P. developed computational analysis tools. E.R.P. performed the experiments. C.D. and E.R.P. processed and analyzed the experimental data. C.D., E.R.P., and F.N. interpreted the results. E.R.P. and F.N. acquired the funding. F.N. supervised the study. C.D., E.R.P., and F.N. wrote the manuscript.

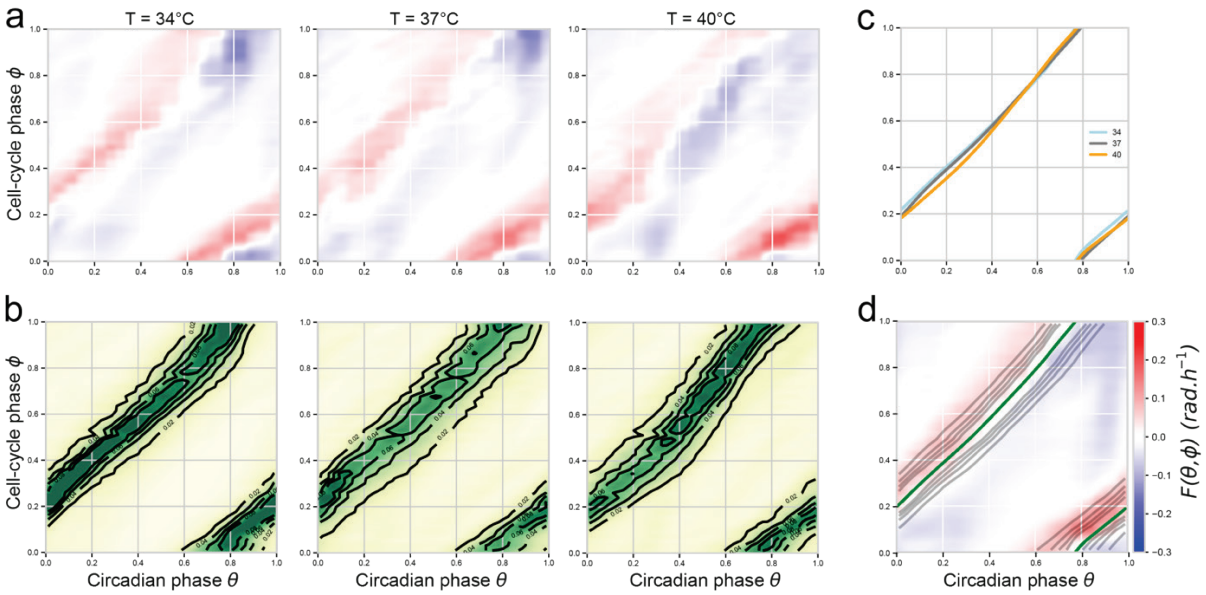
2.10. References

For consistency, references from the article have been placed at the end of the thesis.

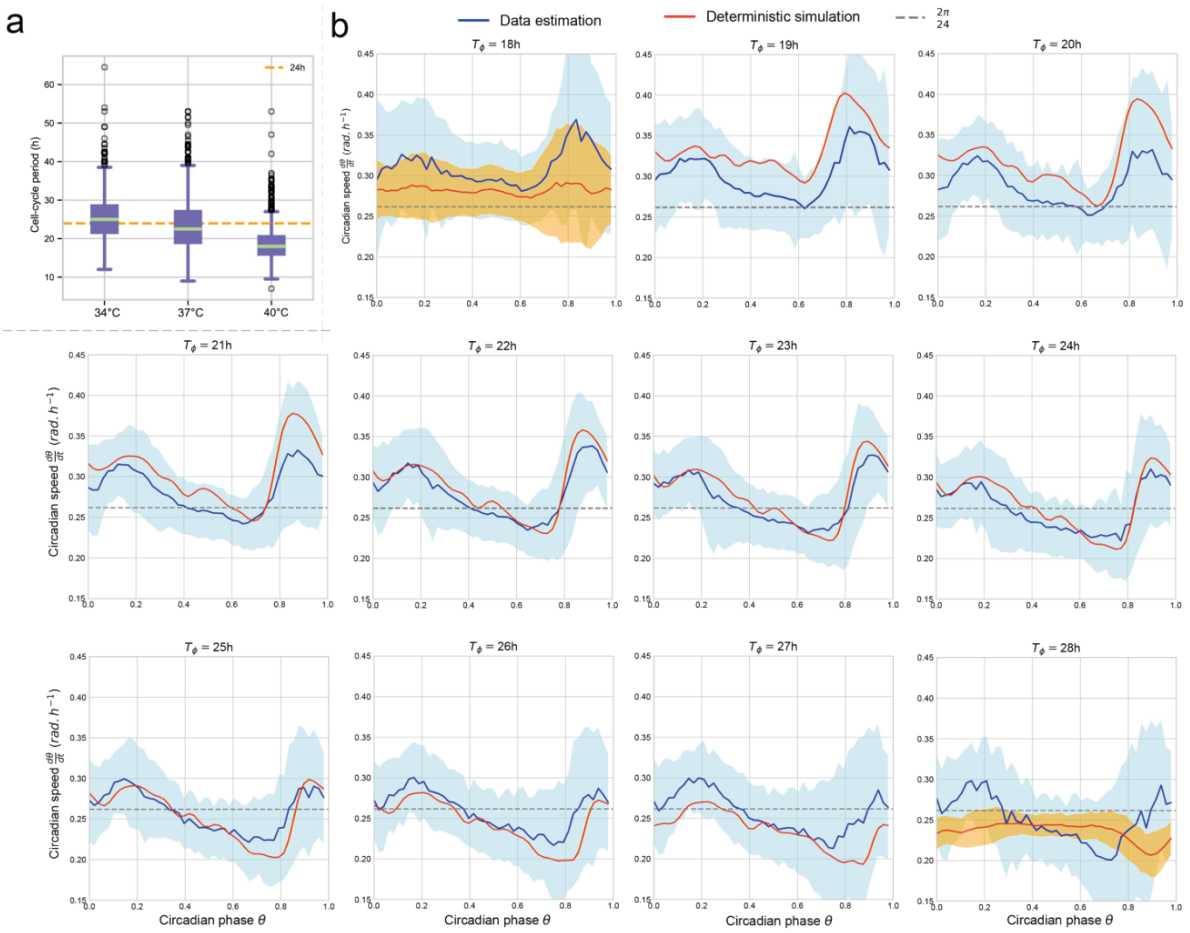
2.11. Supplementary Figures



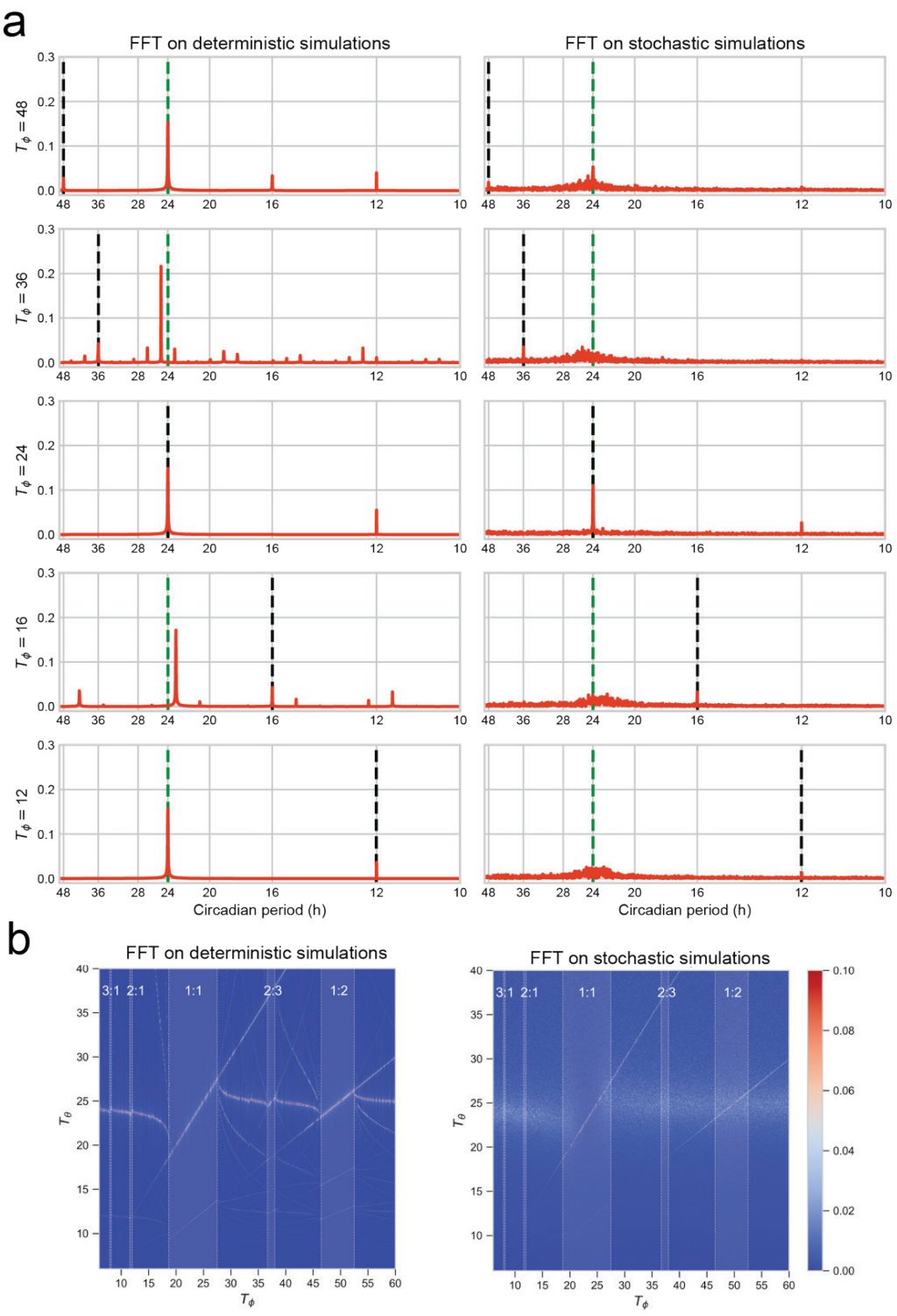
Supplementary Figure 1.1: Validation of the inference methods. (**a**) Examples of temporal data traces and model fits. The colors show data (red), posterior mean for S_t (blue), posteriors circular mean for θ_t (dashed purple), posterior means A_t (dashed grey) and baseline B_t (dashed yellow). The cell-cycle phase ϕ_t (which is not a hidden variable) is obtained from linear interpolation between two successive divisions (dashed green line). Deviations of the purple curve from a straight line corresponds to transient variations of circadian phase velocity, owing to noise and coupling. (**b-c**) Using oscillator parameters mimicking real cells, we can recover coupling functions from simulated traces. (**b**) First, we simulated traces with a coupling $F(\theta,\phi)$ comprising two Gaussian interaction regions, as shown in the left panel. The reconstructed function is shown in the right panel. (**c**) Same numerical experiment made with the coupling inferred from the real data as input (left). Both simulations reveal that the inferred functions are qualitatively accurate, but quantitatively damped.



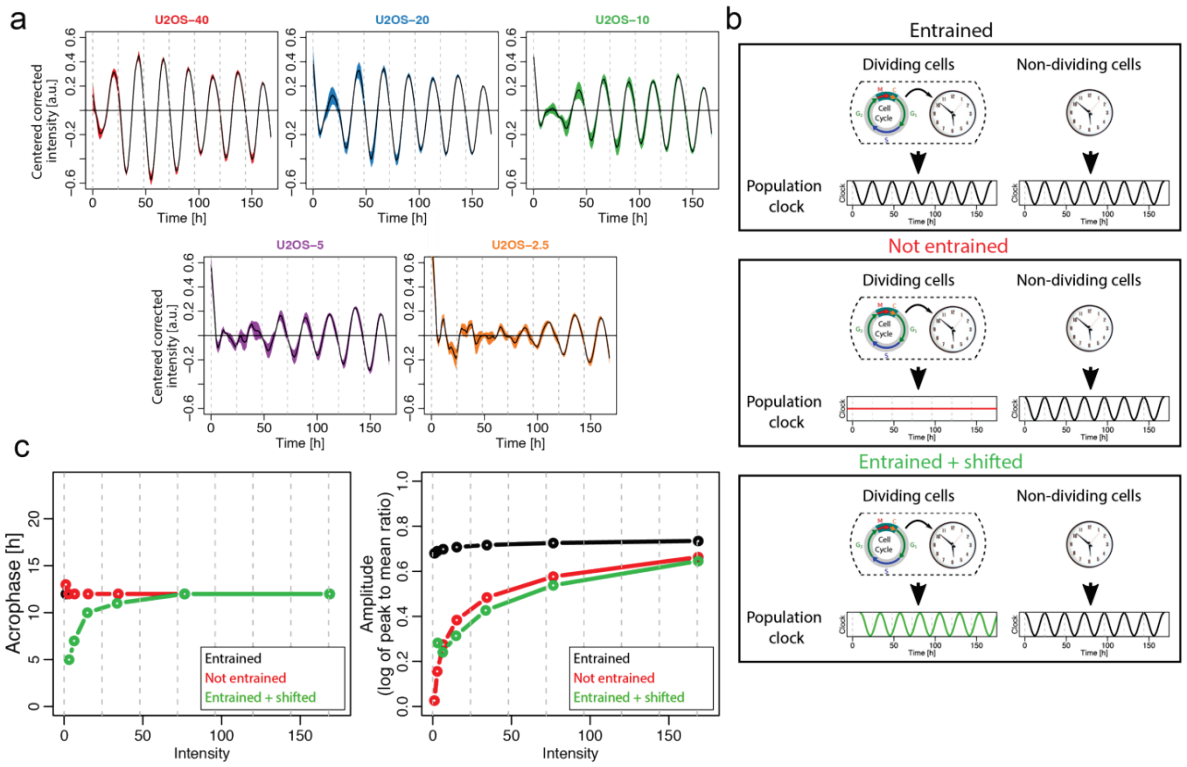
Supplementary Figure 1.2: $F(\theta,\phi)$ depends weakly on temperature. (**a**) Coupling functions obtained from traces acquired at 34°C, 37°C and 40°C. Here, to avoid possible bias, the traces were sampled such that the distribution of cell-cycle periods is identical for each temperature (n = 513 in each case, Supplementary Figure 1.3a for the cell-cycle period distributions at the three different temperatures). (**b**) Superimposition of the merged (from all traces at 34°C, 37°C, and 40°C) coupling function (that of Figure 1.2a), the phase-space trace density (Figure 1.2b, here shown as contour lines), and the attractor (Figure 1.2c, green line). (**c**) Attractors for the vector fields in a) show that the 1:1 phase-locked orbit is temperature independent. (**d**) Phase space densities obtained in the same condition and from the same traces as in (a).



Supplementary Figure 1.3: The phase velocity profiles along the 1:1 attractors in function of cell-cycle period. (**a**) Cell-cycle period (division-to-division time intervals) distributions in NIH3T3 cells at three temperatures. The average period progressively decreases from around 24h at 34°C to 18h at 40°C. (**b**) Circadian phase velocity on the 1:1 attracting orbit in function of cell-cycle period (increasing from left to right and top to bottom). The phase velocity shown is either inferred from the data traces (blue line, standard deviation in light blue), or simulated using the deterministic model (no phase noise) (orange line). The first and last panels (T_ϕ = 18h and 28h) have quasi-periodic orbits (hence a standard deviation is associated with the mean phase velocity). The natural (non-dividing cells) circadian phase velocity (about 0.26 rad.h^{-1}, corresponding to a 24h period) is indicated by a dashed grey line.



Supplementary Figure 1.4: Spectral analysis of simulated traces shows that 1:1 phase-locking is robust against noise. (**a**) Spectral analysis of long simulated circadian traces (t_f=10.000h) using either the deterministic (left, phase diffusion set to zero) or stochastic (right) model, for different cell-cycle periods. Periods of the natural circadian period (24h, green dashed line), and that of the entraining cell cycle (black dashed lines) are indicated. (**b**) Power spectra presented in (a) shown as heatmaps for 350 different cell-cycle periods (see also Movie 1.3). Phase-locked intervals are observed in the deterministic model (left) as lines for the fundamental and few harmonics. Only 1:1 and the 1:2 are visible in the presence of noise (right).



Supplementary Figure 1.5: The circadian clock of dividing cells does not entrain to temperature cycles. (**a**) Averaged (n=6) Bmal1-Luc intensities and 95% confidence intervals from U2OS-Dual Bmal1 luciferase signal centered and corrected for temperature artifact (Supplementary Information). Results were obtained by plating different number of cells (40k, 20k, 10k, 5k, or 2.5k) at the beginning of the experiment. (**b**) Pictograms depicting three different models: i) the circadian oscillators in dividing cells adopt the same circadian profile as non-dividing entrained cells; ii) are not entrained; or iii) are entrained, but with a different phase compared to non-dividing cells. (**c**) Acrophase (hour of the peak of the signal) (left) and amplitude (log of peak to mean ratio) (right) in function of intensity obtained from simulations of the three models in b). Results here should be compared with **Figure 1.6**b in the main text.

2.12. Supplementary Table and Movies

Supplementary Table 1.1 and Movies 1.1-3 are not adapted to the format of this thesis but are available online along with both the published and preprint version of this study.

2.13. Supplementary information

2.13.1. Reconstruction of the dynamical model

The main objective is to perform a data-driven reconstruction of a stochastic models for the coupled systems of circadian and cell cycles, and to then analyze the consequences on the coupled oscillator dynamics. A key step is to estimate the coupling function $F(\theta, \phi)$

(expressed in terms of the phases of the two oscillators) representing the influence of the cell-cycle on the circadian clock.

Our approach consists in explicitly modeling the measured fluorescent signals, using a set of stochastic ordinary differential equations (SODEs) whose parameters are estimated by maximizing the probability of observing the data over the entire set of cell traces. Parameters of the SODEs, which include the oscillator coupling, are assumed to be shared by all cells of a given experimental condition. The method uses several steps, which are detailed in the following sections.

2.13.1.1. Stochastic models for the oscillator phases and measured signals

2.13.1.1.1. Phase model for dividing cells

The circadian phase, representing the state of the circadian oscillator of an individual cell, is modeled as a diffusion-drift process, while the cell-cycle has simpler, piece-wise linear dynamics between two divisions. This is motivated by previous work where we have shown that the influence of the clock on the cell-cycle was very weak, and probably nonexistent[45]. Therefore we focus here on a precise characterization of the coupling function representing the cell-cycle influence on the clock, and then study the dynamical implications.

We first introduce some notation. $\theta, \phi \in [0; 2\pi[\times [0; 2\pi[$ represent the phases of the circadian clock and the cell-cycle, respectively. The intrinsic period of the circadian clock, T_θ, is kept fixed to $24h$, while the cell-cycle intervals T_ϕ^i are indexed on the division-to-division interval i. The coupling function $F(\theta, \phi)$ represents the influence of the cell-cycle phase on the circadian clock phase. σ_θ is the noise strength of the circadian phase, the noise itself being modelled through a Wiener process $W_{\theta,t}$. The stochastic phase model is a two-dimensional diffusion drift written as follows:

$$\begin{cases} \mathrm{d}\theta_t = \dfrac{2\pi}{T_\theta}\mathrm{d}t + F(\theta_t, \phi_t)\mathrm{d}t + \sigma_\theta \, \mathrm{d}W_{\theta,t} \\ \mathrm{d}\phi_t = \dfrac{2\pi}{T_\phi^i}\mathrm{d}t \end{cases} \quad (1)$$

2.13.1.1.2. Phase model for non-dividing cells

Due to inherent fluctuations in biological processes, there are always cells in a dish which transiently exit the cell cycle. On the other hand, the circadian cycle proceeds unperturbed also in quiescent cells. For such cell traces without division, we assume that the model for the circadian phase reduces to just one stochastic ordinary differential equation:

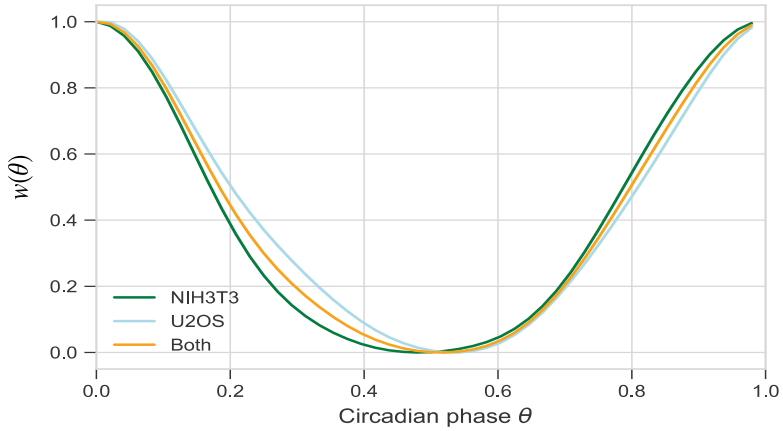
$$\mathrm{d}\theta_t = \frac{2\pi}{T_\theta}\mathrm{d}t + \sigma_\theta \mathrm{d}W_{\theta,t} \qquad (2)$$

containing the parameters of the bare oscillator. In fact, we will use such traces to estimate those parameters.

2.13.1.1.3. Model for the fluorescence signals

The experimental signals obtained from microscopy show noisy oscillations with variations in the amplitude of the maxima as well as in the fluorescence background. For convenience, we centered and rescaled all single cell traces such that the 5th percentile is 0 and 95th percentile is 1.

The phase θ_t is linked to the signal S_t via a function $w(\theta_t)$, which thus defines the phase in our model. In order to use a common definition of the phase, in particular one that does not depend on temperature or cell type (*i.e.* NIH3T3 and U2OS cells), we estimated a single function $w(\theta_t)$, as the the average of all peak-to-peak signals of non-dividing cells. This showed that indeed NIH3T3 and U2OS cells yield very similar functions, and we therefore used the average as a fixed function for all analyses (Supplementary Information Figure 1.1).



Supplementary Information Figure 1.1: Estimated $w(\theta)$ from NIH3T3 (green) and U2OS (blue) traces reveal little difference between them. To keep a consistent definition for the phase, the final function was taken as the average (yellow).

To take into account the variations of amplitude and background, two Ornstein-Uhlenbeck (O-U) processes are used: A_t and B_t, modelled as stochastic ordinary differential equations:

$$\begin{cases} \mathrm{d}A_t = -\gamma_A(A_t - \mu_A)\mathrm{d}t + \sigma_A\,\mathrm{d}W_{A,t} \\ \mathrm{d}B_t = -\gamma_B(B_t - \mu_B)\mathrm{d}t + \sigma_B\,\mathrm{d}W_{B,t} \\ S_t = \exp(A_t)w(\theta_t) + B_t + \xi \end{cases} \qquad (3)$$

In this parametrization, the stationary mean and variance are $\mathrm{E}[X_t] = \mu_X$ and $\mathrm{Var}[X_t] = \frac{\sigma_X^2}{2\gamma_X}$, for $X = A, B$, respectively. ξ represents additional experimental (measurement) white noise with zero mean and variance σ_e^2.

This model assumes that the amplitude and the background fluctuations are independent from the phase (see 2.13.1.1.4 in this document).

2.13.1.1.4. Model conversion into a Hidden Markov Model (HMM)

A HMM is defined as a stochastic triplet $\mathbf{\Omega} = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{E}\}$, where $\boldsymbol{\pi}$ is the vector containing the initial probability distribution of the modeled Markov process, and the matrices A and E contain the transition and emission probabilities of the process[141].

To define the transition and emission matrices, we first discretize the model. The discrete phase, amplitude and background domains are defined as:

$$\begin{cases} \mathbf{\Psi} = \{k\Delta_\psi | k \in \{0,1,\dots,N-1\}, \Delta_\psi = 2\pi/N\} = \{\psi_0,\dots,\psi_{N-1}\} \\ \mathcal{A} = \{A_{min} + k\Delta_A | k \in \{0,1,\dots,M-1\}, \Delta_A = \frac{A_{max}-A_{min}}{M}\} = \{a_0,\dots,a_{M-1}\} \\ \mathcal{B} = \{B_{min} + k\Delta_B | k \in \{0,1,\dots,M-1\}, \Delta_B = \frac{B_{max}-B_{min}}{M}\} = \{b_0,\dots,b_{M-1}\} \end{cases} \quad (4)$$

with N and M the numbers of hidden states for the phase and the O-U processes, respectively.

The maxima (A_{max}, B_{max}) and minima (A_{min}, B_{min}) are chosen at least three standard deviations away from the mean of the corresponding O-U processes. The full hidden state space is then $\mathcal{X} = \mathbf{\Psi} \times \mathcal{A} \times \mathcal{B}$. The transition probabilities are then obtained from the following:

$$p(\theta_{t+dt} | \theta_t = \psi_i, \phi_t = \psi_j) = N\left(\psi_i + \frac{2\pi}{T_\theta}dt + F(\psi_i, \psi_j)dt, \sigma_\theta^2 dt\right) \quad (5)$$

where we have made the approximation that dt is small.

For the O-U processes, the results are well-known[166]:

$$p(A_{t+dt} | A_t = a_i) = N\left(\mu_A + (a_i - \mu_A)e^{-\gamma_A dt}, (1 - e^{-2\gamma_A dt})\frac{\sigma_A^2}{2\gamma_A}\right) \quad (6)$$

$$p(B_{t+dt} | B_t = b_i) = N\left(\mu_B + (b_i - \mu_B)e^{-\gamma_B dt}, (1 - e^{-2\gamma_B dt})\frac{\sigma_B^2}{2\gamma_B}\right) \quad (7)$$

All the transitions between the three processes are assumed to be independent:

$$\mathrm{A}_{ijk,lm,no} = p_{tr}(\psi_k | \psi_i, \psi_j) P_{tr}(a_m | a_l) P_{tr}(b_o | b_n) \quad (8)$$

Finally, the probability of an observation O_t obeys to:

$$p\left(O_t \middle| A_t = a_i, B_t = b_j, \theta_t = \psi_k\right) = \frac{1}{\sigma_e\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\exp(a_i)w(\psi_k)+b_j-O_t}{\sigma_e}\right)^2} \quad (9)$$

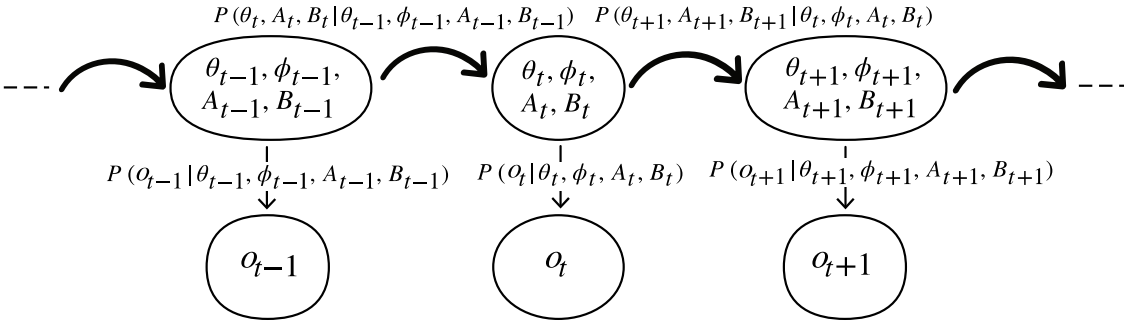
Note that w appears in this equation, enabling to compare the prediction of the model with the actual observations. Also note that in this SI we use O_t to refer to a specific observation in our dataset, instead of S_t in the generic case. However, both notations are interchangeable.

The emission matrix $\mathbf{E_t}$ for each time point can then be computed as:

$$\mathbf{E}_{t,ijk} = p_e\left(O_t \middle| a_i, b_j, \psi_k\right) \quad (10)$$

The fixed cell-cycle phases, given at each time point through a linear interpolation are noted as $\mathbf{\Phi}$, such that $\mathbf{\Phi} = \{\phi_0, \dots, \phi_t, \dots, \phi_T\}$ with $\phi_t \in \mathbf{\Psi} \forall t$.

Once the matrices $\boldsymbol{\pi}, \mathbf{A}$ and $\mathbf{E}$ are built, one can compute the probability of a state $\mathbf{x} \in \mathcal{X}$ given all the observations $\mathrm{O} = \{O_1, \dots, O_t, \dots, O_T\}$, that is the posterior state distribution $p(\mathbf{x}|\mathrm{O})$, using the forward-backward algorithm[141]. A graphical representation of the HMM is is provided Supplementary Information Figure 1.2.



Supplementary Information Figure 1.2: Representation of the Hidden Markov Model used for phase inference, at time $\mathrm{t}-1$, t and $\mathrm{t}+1$. Bold arrows correspond to state transitions, while light arrow correspond to state emissions.

2.13.1.2. Parameters of the model

In all analyses, the number of states for the phase, N, and for the O-U processes, M, were taken as $N = 48$ and $M = 30$, yielding a total number of discrete states of 43200.

2.13.1.2.1. Parameters common to dividing and non-dividing cells

We first discuss the parameters describing the circadian oscillations in individual cells, which we assume as independent from the coupling with the cell cycle. These parameters concern the oscillator period, phase noise, amplitude and background processes, as well as the experimental noise. The parameters were estimated as described below, and are given

in Supplementary Information Table 1.1 for both NIH3T3 and U2OS cells. Since the estimates were found to be very similar at the three experimental temperatures, we considered fixed (temperature-independent) values.

	$T_\theta(h)$	$\sigma_\theta(rad.h^{-1/2})$	μ_A	σ_A	μ_B	σ_B	$\gamma_A(h^{-1})$	$\gamma_B(h^{-1})$	σ_e
NIH3T3	24	0.16	-0.28	0.11	0.08	0.05	0.075	0.075	0.15
U2OS	24	0.18	-0.45	0.15	0.06	0.05	0.075	0.075	0.15

Supplementary Information Table 1.1: Set of parameters for the single cell circadian oscillators in NIH3T3 and U2OS cells. The values of σ_A, σ_B and σ_e are in units of the centered and rescaled signals.

The circadian oscillator period T_θ was estimated by averaging the peak to peak times in Rev-Erbα-YFP signals on the whole set of non-dividing traces. The resulting value was 24.28h (NIH3T3, whole dataset), rounded for convenience.

To estimate the phase noise σ_θ, we used the property that the peak-to-peak time distribution of the circadian phase (modeled as a diffusion-drift process) θ obeys:

$$T_{2\pi} \sim IG\left(\mu = \frac{2\pi}{\omega_\theta} = T_\theta, \lambda = \frac{(2\pi)^2}{\sigma_\theta^2}\right) \qquad (11)$$

where $IG(\lambda, \mu)$ stands for the inverse Gaussian distribution with mean μ and shape parameter λ. This distribution has variance μ^3/λ. Therefore:

$$\sigma_\theta^2 = \frac{\mathrm{Var}[T_{2\pi}]4\pi^2}{T_\theta^3} \qquad (12)$$

This expression was used in the NIH3T3 cells. Because we observed only very few non-dividing U2OS cells, we needed to estimate σ_θ from the dividing traces. Since we observed from traces generated *in silico* that the cell-cycle coupling added about 35% of variability in the peak-to-peak distribution, we corrected the value of σ_θ obtained from dividing U2OS cells for this effect.

The means and noise of the O-U processes were estimated from the set of all minima and maxima of the non-dividing traces. More precisely, the mean background was calculated as the average minimum value of the signal, and the mean log amplitude as the average log difference between the maxima and surrounding minima. Similarly, the noise strengths were obtained from the variances of those quantities, using the relationship for the stationary variances: $\sigma_X^2 = 2\gamma_X^2 \mathrm{Var}[X]$, for $X = A, B$.

We assumed that the time constants of the A and B processes were slower than the phase fluctuations occurring within one oscillatory cycle, and therefore chose $\gamma_A = \gamma_B = 1/14h^{-1}$.

We verified that values of γ_A and γ_B in the range of $1/5h^{-1}$ to $1/30h^{-1}$ did not lead to major differences in the resulting coupling function.

The noise parameter σ_e was set to 0.15. Since the signals were quantile normalized (see Section 2.13.1.1.3), this corresponds to a relative error of about 15%.

2.13.1.2.2. Parameters of the coupling function

2.13.1.2.2.1. The EM algorithm

Here we introduce the expectation-maximization (E-M) algorithm, which will be used to estimate the coupling function.

Denote the sequence of observations by O, the state space by $\mathcal{X}$, a given sequence of states by $\mathbf{X}$ and the current and updated set of parameters by $\mathbf{\Lambda}'$ and $\mathbf{\Lambda}$, respectively. The Q function of the EM[126] is written:

$$Q(\mathbf{\Lambda}, \mathbf{\Lambda}') = \sum_{\mathrm{X} \in \mathcal{X}} \log\big(p(\mathrm{O}, \mathbf{X}|\mathbf{\Lambda})\big) p(\mathbf{X}|\mathrm{O}, \mathbf{\Lambda}') \quad (13)$$

Here, the sequence $\mathbf{X}$ is composed of states x such that $\mathrm{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_t, \ldots, \mathrm{x}_T\}$, and these states are themselves composed of three substates for the phase, the amplitude and the background: $\mathrm{x} = (\psi, a, b)$. In our problem, if we define i_0 as the index associated with the first state of the sequence $\mathbf{X}$, the probability of the observations and the states can be written as a product:

$$p(\mathrm{O}, \mathbf{X}|\mathbf{\Lambda}) = \boldsymbol{\pi}_{i_0} \prod_{t=1}^{T} p\left(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{\Lambda}\right) p(O_{t+1}|\mathbf{X}_{t+1}, \mathbf{\Lambda}) \quad (14)$$

This equation enables to rewrite the function Q as three separated sums:

$$\begin{aligned} Q(\mathbf{\Lambda}, \mathbf{\Lambda}') = & \sum_{\mathrm{X} \in \mathcal{X}} \log\left(\boldsymbol{\pi}_{i_0}\right) p(\mathbf{X}|\mathrm{O}, \mathbf{\Lambda}') + \sum_{\mathrm{X} \in \mathcal{X}} \left(\sum_{t=1}^{T} \log\left(p(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{\Lambda})\right) \right) p(\mathbf{X}|\mathrm{O}, \mathbf{\Lambda}') \\ & + \sum_{\mathrm{X} \in \mathcal{X}} \left(\sum_{t=1}^{T} \log\left(p(O_{t+1}|\mathbf{X}_{t+1}, \mathbf{\Lambda})\right) \right) p(\mathbf{X}|\mathrm{O}, \mathbf{\Lambda}') \quad (15) \end{aligned}$$

This expression readily extends to several traces by adding another sum over the trace indices. Each term can now be optimized individually, enabling to find the optimal set of parameters for the initial condition, the state transitions (which contain the coupling function), and the emissions.

2.13.1.2.2.2. Estimation of the initial condition

Taking the derivative of the first term in Eq (15) with respect to the components of $\boldsymbol{\pi}$ leads to the optimal initial condition:

$$\boldsymbol{\pi}_i = p(\mathbf{X}_0 = \mathrm{x}_i|\mathrm{O}, \boldsymbol{\Lambda}') \qquad (16)$$

2.13.1.2.2.3. Estimation of the coupling function

The coupling function is parameterized on a grid of N^2 parameters, such that F_{ij} corresponds to the coupling for the pair of phases $(\theta_i, \phi_j) \in \boldsymbol{\Psi}^2$. Due to this high number of parameters, regularization constraints were added. Specifically, the squared norm of the gradient, $||\nabla F_{ij}||^2 = (\frac{F_{i+1,j}-F_{i,j}}{\Delta\psi})^2 + (\frac{F_{i,j+1}-F_{i,j}}{\Delta\psi})^2$ is used to control for smoothness. In addition, we controlled the sparseness of the coupling function using the squared norm. The penalized version of Q is therefore:

$$Q_p(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') = Q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') - \lambda_1 \sum_{i,j} |\,||\nabla F_{ij}||^2 - \lambda_2 \sum_{i,j} {F_{ij}}^2 \qquad (17)$$

Starting again from Eq. (17) augmented with these new penalization terms yields:

$$\frac{\partial Q_p(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}')}{\partial F_{kl}} = \overbrace{\frac{\partial}{\partial F_{kl}}\left[\sum_{\mathrm{X}\in\mathcal{X}}\left(\sum_t \log\left(p(\mathbf{X}_{t+1}|\mathbf{X}_t, \boldsymbol{\Lambda})\right)\right)p(\mathbf{X}|\mathrm{O}, \boldsymbol{\Lambda}')\right]}^{E_1} - \overbrace{\frac{\partial}{\partial F_{kl}}\left[\lambda_1 \sum_{i,j} |\,|\nabla F_{ij}||^2 + \lambda_2 \sum_{i,j} F_{ij}^2\right]}^{E_2} \quad (18)$$

The first part of this equation, E_1, corresponds to the state transitions, while the second part, E_2, corresponds to the penalization. Equating this to zero to find the maxima conditions, and explicitly taking the sequence of cell-cycle states into account, we obtain:

$$\frac{\partial}{\partial F_{kl}}\Big[\sum_{i_1,i_2}\sum_{j_1,j_2}\sum_{k_1,k_2}\sum_t \log\left(p(\theta_{i_2}, a_{j_2}, b_{k_2}|\theta_{i_1}, \phi_t, a_{j_1}, b_{k_1}, \boldsymbol{\Lambda})\right) p(\mathrm{x}_t = (\theta_{i_1}, a_{j_1}, b_{k_1}), \mathrm{x}_{t+1} = (\theta_{i_2}, a_{j_2}, b_{k_2})|\mathrm{O}, \boldsymbol{\Lambda}')\Big] = E_2 \qquad (19)$$

Note here that $\theta_{i_1}, \theta_{i_2}, a_{j_1}, a_{j_2}, b_{k_1}, b_{k_2}$ are hidden states, for which we infer a distribution of probability with the HMM, while ϕ_t is given as an external parameter. Now, the Markov propagators for the phase, amplitude and background being independent (cf. Eq. 8), we have:

$$\begin{aligned}\log\big(p(\theta_{i_2}, a_{j_2}, b_{k_2}|\theta_{i_1}, \phi_t, a_{j_1}, b_{k_1}, \boldsymbol{\Lambda})\big) &= \log\big(p(\theta_{i_2}|\theta_{i_1}, \phi_t, \boldsymbol{\Lambda})\big) \\ &+\log\big(p(a_{j_2}|a_{j_1}, \boldsymbol{\Lambda})\big) + \log\big(p(b_{k_2}|b_{k_1}, \boldsymbol{\Lambda})\big)\end{aligned} \qquad (20)$$

Since the transitions probabilities for the amplitude and the background do not depend on the coupling function, they cancel out with the derivative. The remaining sum leads to the

marginal joint distribution of phases at time t and $t+1$. To keep continuity with the previous notation, we denote the marginal $p(\theta_t = \theta_i) = \sum_{\theta_i, a_j, b_k} p\,(\mathrm{x}_t = (\theta_i, a_j, b_k))$ by $p(\mathrm{x}_t = (\theta_i, ., .))$.

We therefore have:

$$\frac{\partial}{\partial F_{kl}}\left[\sum_{i_1,i_2}\sum_{t}\log\left(p\left(\theta_{i_2}\middle|\theta_{i_1},\phi_t,\boldsymbol{\Lambda}\right)\right)p\left(\mathrm{x}_t=\left(\theta_{i_1},.,.\right),\mathrm{x}_{t+1}=\left(\theta_{i_2},.,.\right)\middle|\mathrm{O},\boldsymbol{\Lambda}'\right)\right]=E_2 \quad (21)$$

Now, $p(\mathrm{x}_t = (\theta_{i_1}, ., .), \mathrm{x}_{t+1} = (\theta_{i_2}, ., .)|\mathrm{O}, \boldsymbol{\Lambda}')$ doesn't depend on the new coupling function parameters F_{kl}, so it can be treated as a multiplicative constant. Defining $\omega_\theta = 2\pi/T_\theta$, this yields:

$$\frac{\partial}{\partial F_{kl}}[\sum_{i_1,i_2}\sum_{t}\quad \log\left(\frac{1}{\sigma_\theta\sqrt{2\pi dt}}e^{-\frac{1}{2}\left(\frac{\theta_{i_2}-(\theta_{i_1}+\omega_\theta dt+F(\theta_{i_1},\phi_t)dt)}{\sigma_\theta^2 dt}\right)^2}\right)$$
$$p(\mathrm{x}_t=(\theta_{i_1},.,.),\mathrm{x}_{t+1}=(\theta_{i_2},.,.)|\mathrm{O},\boldsymbol{\Lambda}')]=E_2 \quad (22)$$

All the terms that do not depend on $F_{kl} = F(\theta_k, \phi_l)$ are removed by the derivative, which simplifies to:

$$\sum_{i_2}\sum_{\{t|\phi_t=\phi_l\}}\frac{\theta_{i_2}-(\theta_k+\omega_\theta dt+F_{kl}dt)}{\sigma_\theta^2}p\left(\mathrm{x}_t=(\theta_k,.,.),\mathrm{x}_{t+1}=\left(\theta_{i_2},.,.\right)\middle|\mathrm{O},\boldsymbol{\Lambda}'\right)$$
$$=E_2 \quad (23)$$

F_{kl} can now be isolated, and we can sum over θ_{i_2} in the denominator:

$$F_{kl}=\frac{-\sigma_\theta^2 E_2+\sum_{i_2}\sum_{\{t|\phi_t=\phi_l\}}\left(\theta_{i_2}-(\theta_k+\omega_\theta dt)\right)p(\mathrm{x}_t=(\theta_k,.,.),\mathrm{x}_{t+1}=(\theta_{i_2},.,.)|\mathrm{O},\boldsymbol{\Lambda}')}{dt\sum_{\{t|\phi_t=\phi_l\}}p\,(\mathrm{x}_t=(\theta_k,.,.)|\mathrm{O},\boldsymbol{\Lambda}')} \quad (24)$$

Note that the function $w(\theta)$ is involved non-explicitly in this equation, through the computation of the posterior phase distributions. From Eq. 18, we find:

$$E_2=\frac{\partial}{\partial F_{kl}}\left[\lambda_1\sum_{i,j}(\frac{F_{i+1,j}-F_{i,j}}{\Delta\psi})^2+(\frac{F_{i,j+1}-F_{i,j}}{\Delta\psi})^2+\lambda_2\sum_{i,j}F_{ij}^2\right] \quad (25)$$

Taking the derivative, and re-injecting into Eq. (24) yields:

$$\begin{aligned} F_{kl} & \left[dt \sum_{\{t|\phi_t=\phi_l\}} p\left(\mathrm{x}_t = (\theta_k,.,.)|\mathrm{O},\boldsymbol{\Lambda}'\right) + \frac{8\lambda_1\sigma_\theta^2}{\Delta\psi^2} + 2\lambda_2\sigma_\theta^2 \right] \\ - & \frac{2\lambda_1\sigma_\theta^2}{\Delta\psi^2} F_{k-1,l} - \frac{2\lambda_1\sigma_\theta^2}{\Delta\psi^2} F_{k+1,l} - \frac{2\lambda_1\sigma_\theta^2}{\Delta\psi^2} F_{k,l+1} - \frac{2\lambda_1\sigma_\theta^2}{\Delta\psi^2} F_{k,l-1} \\ & = \sum_{i_2} \sum_{\{t|\phi_t=\phi_l\}} \left(\theta_{i_2} - (\theta_k + \omega_\theta dt)\right) p(\mathrm{x}_t = (\theta_k,.,.), \mathrm{x}_{t+1} = (\theta_{i_2},.,.)|\mathrm{O},\boldsymbol{\Lambda}') \end{aligned} \quad (26)$$

For readability, we define the new following quantities:

$$\begin{cases} Q_1 = dt \sum_{\{t|\phi_t=\phi_l\}} p\left(\mathrm{x}_t = (\theta_k,.,.)|\mathrm{O},\boldsymbol{\Lambda}'\right) + \frac{8\lambda_1\sigma_\theta^2}{\Delta\psi^2} + 2\lambda_2\sigma_\theta^2 \\ Q_2 = -\frac{2\lambda_1\sigma_\theta^2}{\Delta\psi^2} \\ Q_{k,l} = \sum_{i_2} \sum_{\{t|\phi_t=\phi_l\}} \left(\theta_{i_2} - (\theta_k + \omega_\theta dt)\right) p(\mathrm{x}_t = (\theta_k,.,.), \mathrm{x}_{t+1} = (\theta_{i_2},.,.)|\mathrm{O},\boldsymbol{\Lambda}') \end{cases} \quad (27)$$

This gives:

$$F_{kl}Q_1 + \left(F_{k-1,l} + F_{k+1,l} + F_{k,l-1} + F_{k,l+1}\right)Q_2 = Q_{k,l} \quad (28)$$

This is a linear equation for F_{kl}. Since Eq. (28) holds $\forall k,l \in \mathbb{N}^2$, this can be rewritten as:

$$\mathrm{Ax} = \mathrm{b} \quad (29)$$

Where **A** is a matrix containing the Q_1 and Q_2 terms, **x** is the vector containing the F_{kl} terms and **b** the vector containing the $Q_{k,l}$ terms. Due to the regularization, Q_1 is always invertible.

2.13.1.2.2.4. Regularization

λ_1 is found using four-fold cross-validation, *i.e.* by splitting the NIH3T3 dataset into four chunks and scanning which λ_1 value gives the best generalization, *i.e.* maximizes the likelihood of the left-out test traces. The resulting value is 10^{-6}.

The value of λ_2 is set according to the following principle. The update expression for the coupling function (when $\lambda_1 = 0$) reads:

$$F_{kl} = \frac{\sum_{i_2} \sum_{\{t|\phi_t=\phi_l\}} \left(\theta_{i_2} - (\theta_k + \omega_\theta dt)\right) p(\mathrm{x}_t = (\theta_k,.,.), \mathrm{x}_{t+1} = (\theta_{i_2},.,.)|\mathrm{O},\boldsymbol{\Lambda}')}{2\sigma_\theta^2\lambda_2 + dt \sum_{\{t|\phi_t=\phi_l\}} p\left(\mathrm{x}_t = (\theta_k,.,.)|\mathrm{O},\boldsymbol{\Lambda}'\right)} \quad (30)$$

Thus, λ_2 buffers the sum $dt \sum_{\{t|\phi_t=\phi_l\}} p\left(\mathrm{x}_t = (\theta_k,.,.)|\mathrm{O},\boldsymbol{\Lambda}'\right)$, especially when the latter is small, *i.e.* for the phase-space points which are rarely visited by the cells. Defining T as the total number of time measurements (from all cells), we set:

$$\lambda_2 = \frac{T\lambda_2{}'}{2\sigma_\theta^2}dt. \qquad (31)$$

The interpretation is as follow: given a phase-space state that is visited once in T time points, if $\lambda_2{}' = \frac{1}{T}$ then the corresponding coupling parameter is halved. More visited states lead to more robust coupling parameters, and conversely for less visited states.

In practice, we want to be able to interpret $\lambda_2{}'$ independently of the total number of time points, and we therefore compute it in units of cell-cycle periods, such that the coupling parameter of a state visited once every 200 cell-cycles is halved, that is:

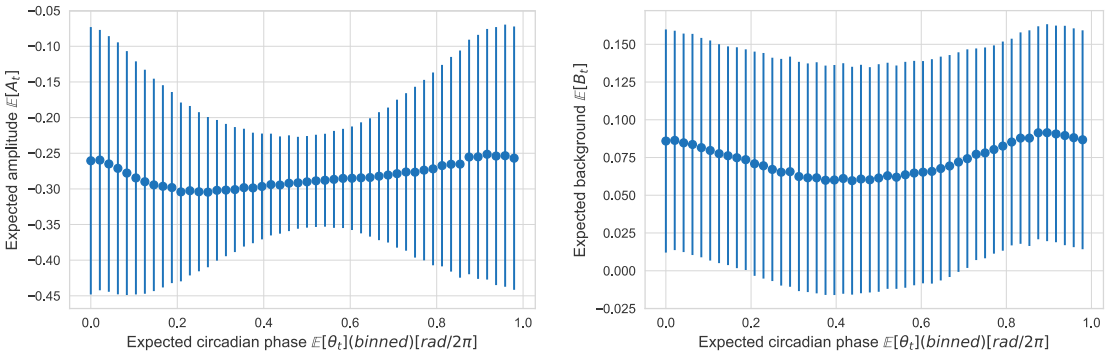
$$\lambda_2' = \frac{1}{200T_\phi} \qquad (32)$$

2.13.1.3. Assessment of model assumptions

In our model for the signal $S_t = \exp(A_t)w(\theta_t) + B_t + \xi$, we assumed that the dynamics of the amplitude, background and phase variables A_t, B_t and θ_t were uncoupled. Within our probabilistic framework, we can *a posteriori* verify this hypothesis, by analyzing the joint posterior distribution $P(\theta_t, A_t, B_t|\mathbf{O})$. Indeed, we can compute the expected values of the three latent variables corresponding to each measured observation O_t as follows:

$$\begin{cases} \mathbb{E}[\theta_t] = \arg(\sum_{i,j,k} p\,(\theta_t = \theta_i, A_t = a_j, B_t = b_k|\mathrm{O})e^{i\theta_i}) \\ \mathbb{E}[A_t] = \sum_{i,j,k} p\,(\theta_t = \theta_i, A_t = a_j, B_t = b_k|\mathrm{O})a_j \\ \mathbb{E}[B_t] = \sum_{i,j,k} p\,(\theta_t = \theta_i, A_t = a_j, B_t = b_k|\mathrm{O})b_k \end{cases} \qquad (33)$$

As shown in Supplementary Information Figure 1.3, both the expected amplitudes and backgrounds are on average only weakly dependent on the expected phases. Indeed, the means for A and B vary by, respectively, less than ±10% and ±20% compared to the global means. In fact, the variation in the means of these expected values in function of the phases is much lower than the spread observed in the phase bins (corresponding to the many measurements with the same expected phases).



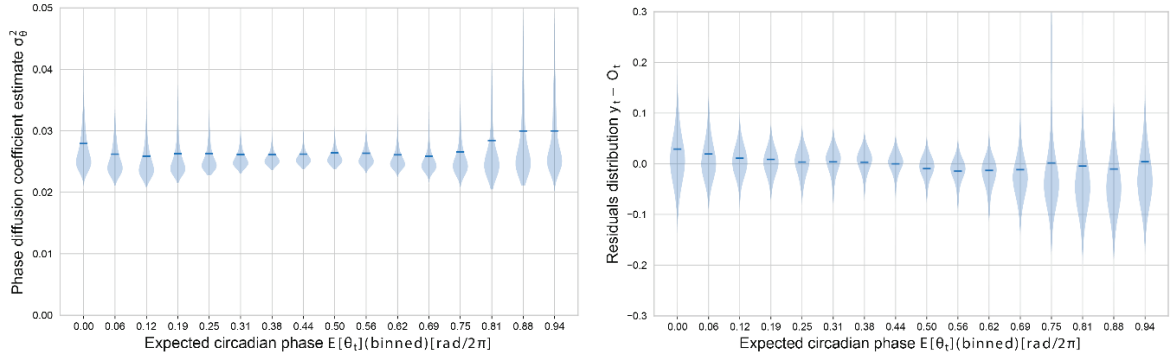
Supplementary Information Figure 1.3: Expected value of the amplitude A_t (**left**) and background B_t (**right**) in function of the expected circadian phase θ_t (binned). The data show the means (dots) and standard deviations computed over all the NIH3T3 cell traces, all temperature included.

Similarly, we then analyzed the (a posteriori) estimated phase noise σ_θ in function of the circadian phase θ_t. To do this, we compute a phase-dependent estimate of σ_θ:

$$\sigma_\theta^2(\theta_i) = \frac{\sum_{j,k,t}\left(\theta_k - (\theta_i + (\omega_\theta + F(\theta_i,\phi_j)dt)\right)^2 p\left(\theta_t = \theta_i, \theta_{t+1} = \theta_k \middle| \phi_t = \phi_j, \mathrm{O}\right)}{dt\sum_{j,k,t} p\left(\theta_t = \theta_i, \theta_{t+1} = \theta_k \middle| \phi_t = \phi_j, \mathrm{O}\right)} \qquad (34)$$

Results (Supplementary Information Figure 1.4, left) reveal only very weak dependence of σ_θ^2 over the circadian phase θ. Indeed, the means in σ_θ vary from 0.016 to 0.017, i.e. a deviation of about ±6%.

Finally, we analyzed the measurement noise ξ. We thus computed the prediction of the model y_t as $y_t = \exp(\mathbb{E}[A_t])w(\mathbb{E}[\theta_t]) + \mathbb{E}[B_t]$, and analyzed the distributions of residuals $y_t - O_t$ binned by expected circadian phase (Supplementary Information Figure 1.4, right). Here, we find that the residuals are centered on 0 within a good approximation (±0.035), showing that there is no systematic bias in the noise model.



Supplementary Information Figure 1.4: Left: Evolution of the distribution of phase diffusion coefficient estimates σ_θ^2 with the expected circadian phase θ_t (binned). **Right:** Evolution of the distribution of residuals $y_t - O_t$ with the expected circadian phase θ_t (binned). Dark horizontal dashes indicate the means. The computations are made from the distributions computed on all the traces coming from NIH3T3 cells, all temperature included.

2.13.1.4. Assessment of the parameter estimation

To assess the parameter estimation, we simulated traces *in silico* and re-estimated the parameters using the same methods as for the experimental traces. The generated traces were of the same scale and length as the experimental traces. The regression parameters γ_A, γ_B and the noise parameters σ_e were taken from Supplementary Information Table 1.1.

Supplementary Information Table 1.2 summarizes the results for all estimated parameters. Although we expect some imprecisions due to the stochasticity of the system, the relative error remains low for every parameter.

	$T_\theta(h)$	$\sigma_\theta(rad.h^{-1/2})$	μ_A	σ_A	μ_B	σ_B
Simulated	24.0	0.16	-0.28	0.11	0.08	0.05
Estimated	24.0	0.16	-0.24	0.11	0.04	0.05

Supplementary Information Table 1.2: Simulated and estimated model parameters.

For the reliability of the estimated coupling function, we refer to the main text (Supplementary Figure 1.1, panels b and c).

2.13.2. Simulations of the dynamical system

2.13.2.1. Model

A deterministic model for the phase dynamics is obtained by removing the phase noise term from the full model. In addition, to study the bifurcations (phase locked states) in function of the coupling strength, we added a multiplicative factor for the coupling function called K: ($K = 1$ for the biological coupling value).

$$\begin{cases} \dot{\theta} = \dfrac{2\pi}{T_\theta} + KF(\theta, \phi) \\ \dot{\phi} = \dfrac{2\pi}{T_\phi} \end{cases} \qquad (35)$$

2.13.2.2. Phase-locked states

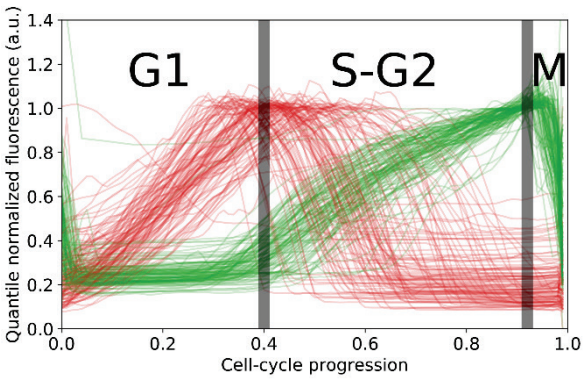
Weakly coupled oscillators can phase-lock when the ratio of their natural period is close to a ratio of integer numbers, *i.e.* $\frac{T_\theta}{T_\phi} \simeq \frac{p}{q}$ with $p, q \in \mathbb{N}$ [1]. To characterize mode-locked states, we estimate $\bar{\omega}_\theta$, defined as the average circadian phase velocity:

$$\bar{\omega}_\theta = \lim_{t \to \infty} \frac{\theta(t)}{t} \qquad (36)$$

Phase-locking occurs when $\bar{\omega}_\theta$ remain constant within an interval of cell-cycle frequencies ω_ϕ, as represented by Arnold tongue diagrams. Outside of such stable intervals, the dynamics is quasiperiodic.

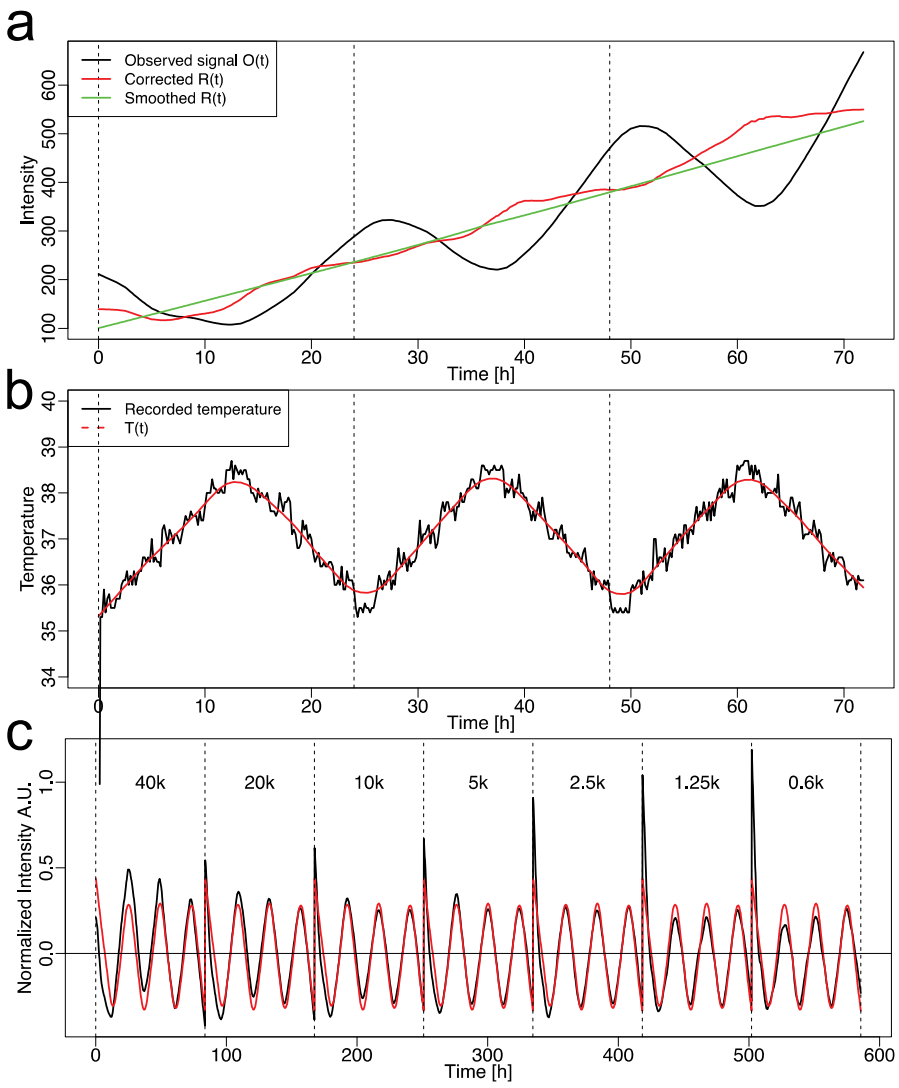
2.13.3. Correspondence between cell-cycle phase and biological cell-cycle events

In our model, we assumed a linear progression of the cell-cycle phase between two successive divisions. To get a better handle on the relation between this measure and cell-cycle events, we generated a set of 104 experimental traces from NIH3T3 cells expressing the FUCCI cell-cycle sensor[123]. To obtain estimates of the boundaries for the different cell-cycle events, we normalized and rescaled all fluorescent signals before mapping them to a 0 to 2π interval (from division to division, Supplementary Information Figure 1.5). Despite biological variability, the growth phase 1 (G1) generally spans from 0 to $0.4 \times 2\pi$ rad, while DNA replication and growth phase 2 (S-G2) usually occur between $0.4 \times 2\pi$ rad and $0.95 \times 2\pi$ rad. Mitosis usually happens from $0.95 \times 2\pi$ rad to 2π rad.



Supplementary Information Figure 1.5: Normalized experimental traces from NIH3T3 cells expressing the FUCCI cell-cycle reporter system enable the association between the physical cell-cycle phase and the biological phase. The red and green fluorescence signals correspond respectively to mKO2-Cdt1 and mAG-Geminin FUCCI reporters. The vertical grey lines denote the (approximate) separation between the different biological cell-cycle phases.

2.13.4. Analysis of a population of bioluminescence traces under temperature entrainment



Supplementary Information Figure 1.6: (**a**) Observed $O(t)$, smoothed $R(t)$ and corrected $R(t)$ signals obtained from U2OS cells expressing a PGK-luciferase reporter grown at low cell confluence. (**b**) Recorded 35.5°C-38.5°C temperature entrainment (black) and smoothed signal (red, $\mathrm{T(t)}$) from (a). (**c**) Normalized PGK-Luc signal obtained from U2OS cells grown at different confluences (black) and the optimal fit using $t_d = 80$ min and $k = -0.26$ (red).

The enzymatic activity of luciferase is known to be higher at lower temperature[164]. Since we applied temperature cycles from 35.5°C to 38.5°C for entrainment, even a luciferase reporter driven by a constitutive gene, *e.g. Pgk*, would show an oscillatory signal (Supplementary Information Figure 1.6, panels a and b)[167], [168]. To correct the signal for this systematic effect, we found that the observed signal $O(t)$ could be well fitted by the following expression:

$$O(t) = R(t)(1 + k(T(t - t_d) - T_0) \quad (37)$$

where $R(t)$ is the real signal exempts of any temperature artifact, $T(t)$ is the temperature profile, $T_0 = 37$°C), $k =$ is a magnitude coefficient, and t_d minutes a time delay. To determine the free parameters t_d and k, we used the luciferase signal obtained from U2OS cells expressing a PGK luciferase reporter (U2OS-PGK-Luc) which is expected to yield a non-oscillating signal after correction (Supplementary Information Figure 1.6a). Specifically, we optimized d and k to best fit $O(t)$, after smoothing $O(t)$ to obtain a proxy for $R(t)$. The optimal fit yielded $t_d = 80$ minutes and $k = -0.26$ (Supplementary Information Figure 1.6c). These values of t_d and k were then used to detrend the circadian luminescence signals using Eq. (37). Importantly, we performed all our luciferase experiments using the Luc2p luciferase (Promega), a destabilized version of the WT Photinus pyralis luciferase optimized for expression in mammals. Consequently, we could use the optimized t_d and k to retrieve the corrected signal $R(t)$ for all our constructs.

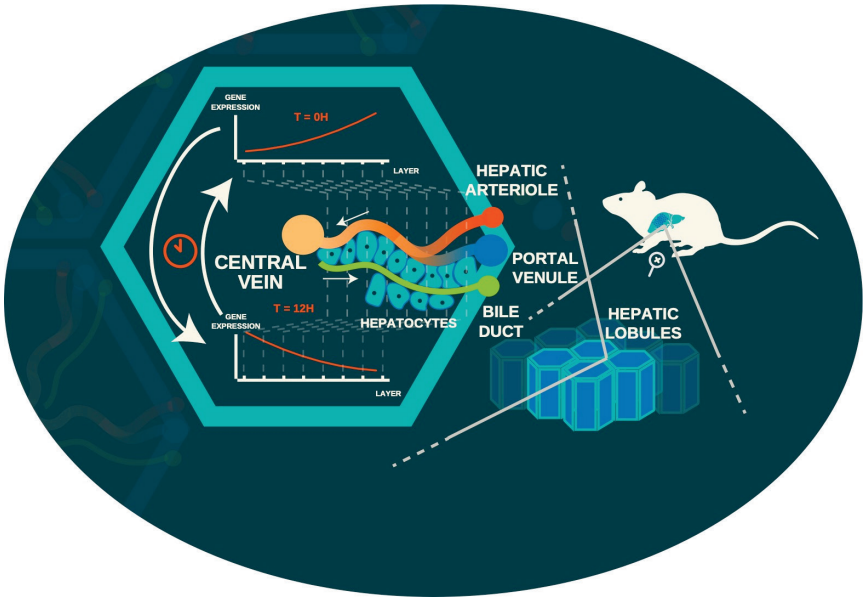
Chapter 2: Space-time logic of liver gene expression at sublobular scale

This work is currently under review at *Nature Metabolism*. It is available as a preprint on BioRxiv (https://doi.org/10.1101/2020.03.05.976571). Authors list is as follow (first authors in bold): **C. Droin, J. El Kholtei, K. B. Halpern**, C. Hurni, M. Rozenberg, S. Muvkadi, S. Itzkovitz, F. Naef.

Contributions

In all the tasks listed below, my work is always under the supervision of F. Naef.

F. Naef and S. Itzkovitz conceived the study. K.B. Halpern, J. El Kholtei and C. Hurni prepared the samples and performed the experiments. K.B. Halpern and C. Hurni did the FISH experiments and parsed the corresponding images (Figure 4, left). J. El Kholtei did the reconstruction and scripts to generate the t-SNE (Figure 1, top). I designed the modelling. I analysed the reconstructed data along with the corresponding fits (Figure 1-4 with corresponding supplementary Figures). Along with J. El Kholtei, I performed the functional analysis (Fig 5-6). With input from S. Itzkovitz and J. El Kholtei, F. Naef and I wrote the full manuscript and I generated/cleaned all Figures and all Tables, including supplementary figures. F. Naef, S. Itzkovitz, J. El Kholtei., K.B. Halpern and I reviewed the manuscript before submission.



Artwork figure 3: Visual abstract of the phenomena in play in circadian zonation: in mouse hepatic lobules, gene expression can vary depending on the positions of the hepatocytes along the portal-central axis (zonation), and on the time of the day (circadian rhythm).

1. Project introduction

1.1. Motivation and aims

Liver gene expression is a fundamental topic in biology and medicine as the liver is crucial for body metabolism. So far, gene expression has been studied either temporally, at the bulk level, either spacially, in specific lobule areas. Temporal studies were made with transcriptomics[149] and proteomics[169]. In contrast, spatial ones were done using either fluorescent in situ hybridisation (FISH)[170], either immunohistochemistry[171], or by FACS-sorting periportal and pericentral hepatocyte populations followed by RNA-seq[172].

Recently, the work of *Bahar Halpern and Shenhav et al.* used single-cell RNA-sequencing to investigate liver zonation on a genome-wide scale with high spatial resolution[173]. This revealed a wide breadth of spatial heterogeneity in mRNA expression that happens to coincide with an intricate organisation of spatially non-uniform liver functions. In parallel, the Naef lab showed how both the circadian clock and the feeding fasting cycles pervasively drive rhythms of gene expression in bulk, impacting key sectors of liver physiology[174].

By extending the work of *Bahar Halpern and Shenhav et al.* to four timepoints distributed along the day, this project aims to provide the first exhaustive analysis of mouse liver gene expression with both spatial and temporal resolution. This should answer how, and to which extent, liver zonation is modulated by circadian rhythms and ultimately how the intricate spatial and temporal activity patterns of the mammalian liver are achieved.

1.2. Background

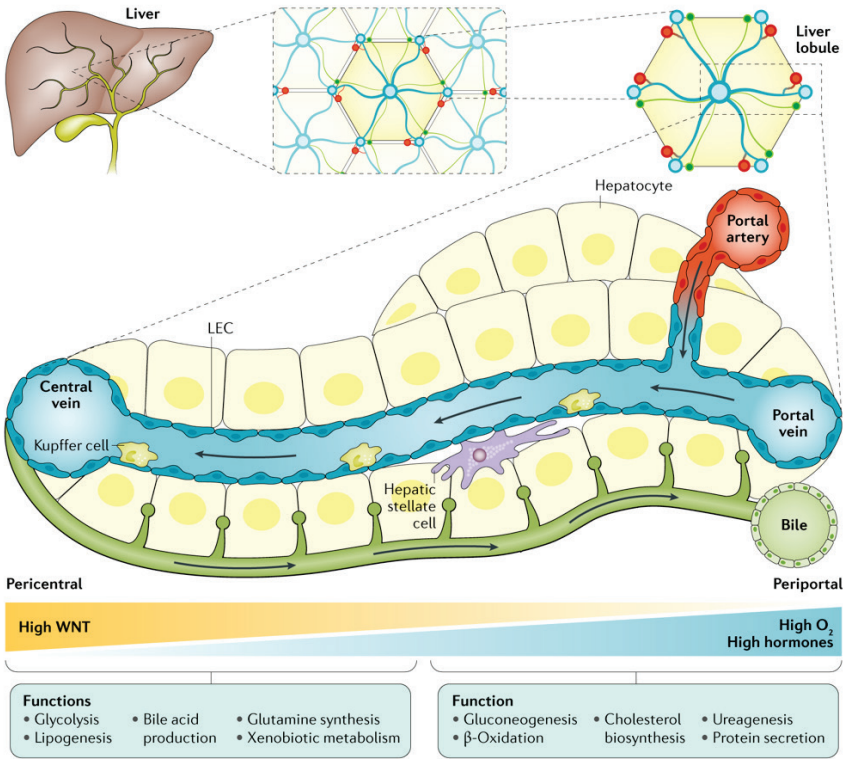
1.2.1. The mammalian liver[22]

The liver is the biggest solid organ of the human body and is crucial for its metabolism. One of its primary functions is regulating the concentration of blood glucose and fatty acids in the blood. It stores glucose and fatty acids in the form of glycogen and triacylglycerides, respectively, and can degrade and release them into the blood on demand. The liver also plays a vital role in processes such as detoxification and bile acid production.

Structurally, the liver is built up of many small (about 1 mm in diameter[175]) hexagonal lobules (Background Figure 2.1). They are the functional unit of the liver and contain portal nodes at the edges and a vein in the centre. The nodes consist of a vein coming from the

[22] Parts of the introduction that follows regarding the mammalian liver is adapted, with permission, from the notes of one colleague and co-author of our study, Jakob El Kholtei.

intestine, an artery coming from the heart and a bile duct. Blood flows from the peripheral artery and vein through tiny blood vessels called sinusoids towards the central vein. This directional blood flow leads to gradients of nutrients and signalling molecules along the sinusoid. The cells of the liver are therefore exposed to different microenvironments defined by their position in the lobule. Research in the last decades has uncovered that the liver's parenchymal cells, the hepatocytes, are influenced by their microenvironment and take up functions specific to the region of the lobule in which they reside. It was shown early that metabolic enzymes such as glutamine synthetase and glucose-6-phosphatase are differentially distributed within a lobule, suggesting that key metabolic processes such as gluconeogenesis and glutamine synthesis are taking place in distinct zones of the liver lobule[176], [177]. This phenomenon is called zonation and, by now, a great number of cellular processes are known to be specifically located in sub-areas of the liver lobule[178], [179]. A recent study from the Itzkovitz lab showed around 50% of expressed liver genes to be spatially zonated[173].



Background Figure 2.1: The liver is composed of hexagonal lobules. Portal triads consisting of a hepatic artery (red), a portal vein (blue) and bile duct (green) are located at the lobule corners, also termed portal nodes. Blood flows through radial sinusoids and drains into the central vein. Concentric layers of hepatocytes are positioned on the axis between the central vein and the portal node. Liver non-parenchymal cells that support hepatocyte function, such as Kupffer cells (light green), liver endothelial cells (LECs; blue and red) and hepatic stellate cells (purple) reside along the lobule axis. Bile secreted from hepatocytes flows from the central to the portal zone through bile canaliculi that drain into the bile duct. Blood flow and secreted morphogens give rise to a spatially graded microenvironment, resulting in different functions assigned to different layers. Figure and caption taken from the review by *Ben-Moshe and Itzkovitz, 2019*[180].

1.2.2. Regulation of liver zonation

Different theories exist as to how liver zonation is achieved and regulated. The observed differences might arise during organismal development or be determined as cells differentiate throughout life[176], [181], [182]. Most evidence, however, supports a model by which the spatial diversity is continuously controlled by concentration gradients of bloodborne nutrients as well as morphogens like wnt and hedgehog[183]. Gradients of bloodborne factors arise due to the blood's entrance to the lobule at the portal nodes and its flow towards the centre, during which the composition of the blood changes as a result of cells taking up and secreting nutrients and other compounds.

The concentration gradient of oxygen is one crucial determinant. Its concentration is highest close to the portal node (periportal), therefore allowing more oxidative phosphorylation to take place in this area[184]. Correspondingly, periportal cells contain more mitochondria than cells close to the central vein (pericentral) and endergonic processes like gluconeogenesis preferentially take place in this area[177], [185], [186]. Direct evidence for the role of oxygen comes from studies showing higher levels of all three hypoxia-inducible factors (HIF-1α, HIF-2α, HIF-3α) as well as erythropoietin (EPO) in the pericentral zone[187], [188]. Following these findings, several hypoxia-activated genes were found to be more expressed close to the central vein, whereas hypoxia-inhibited genes showed the highest expression close to the portal node[173].

Zonation is also strongly influenced by morphogens. Most importantly, wnt/β-catenin signalling was identified as a significant player regulating liver zonation. *Benhamouche et al.* have found APC, which is required for β-catenin degradation, to be present in the portal but not in the central area. In contrast, unphosphorylated (= active) β-catenin was located around the central vein[189]. A liver-specific conditional APC knock-out caused glutamine synthetase, a key marker of the central area, to be expressed in all locations of the lobule. This strongly suggests the role of wnt/β-catenin signalling in liver zonation. Yet the source of activating compounds and mode of action of this pathway is not fully understood. Most of the wnt- and frizzled- (wnt-receptor on the cell surface) genes are expressed in at least one of the liver cell types (hepatocytes, biliary epithelial cells, endothelial cells, stellate or kupffer cells), and several studies have found evidence for intercellular communication influencing wnt/β-catenin signalling[190]. WNT9A secreted from endothelial and stellate cells of the endothelial wall influences hepatocyte proliferation and glycogen accumulation in developing chicks, while *Rspondin3*, an important wnt-activator, is specifically expressed in endothelial cells lining the central vein and its expression is required to maintain liver zonation[191], [192]. An influence of extra-hepatic sources is also conceivable, e.g. via lipoproteins[193].

Other factors such as hedgehog-signalling, hormones and growth factors potentially also contribute to achieving liver zonation[194].

1.2.3. The liver as a circadian organ

While humans are a diurnal (day-active) species, mice are nocturnal, and their intrinsic clock is therefore optimised to prepare their metabolism for physical activity and food intake at night. In mice, nearly 50% of known protein-coding genes are thought to show circadian changes in transcription[149].

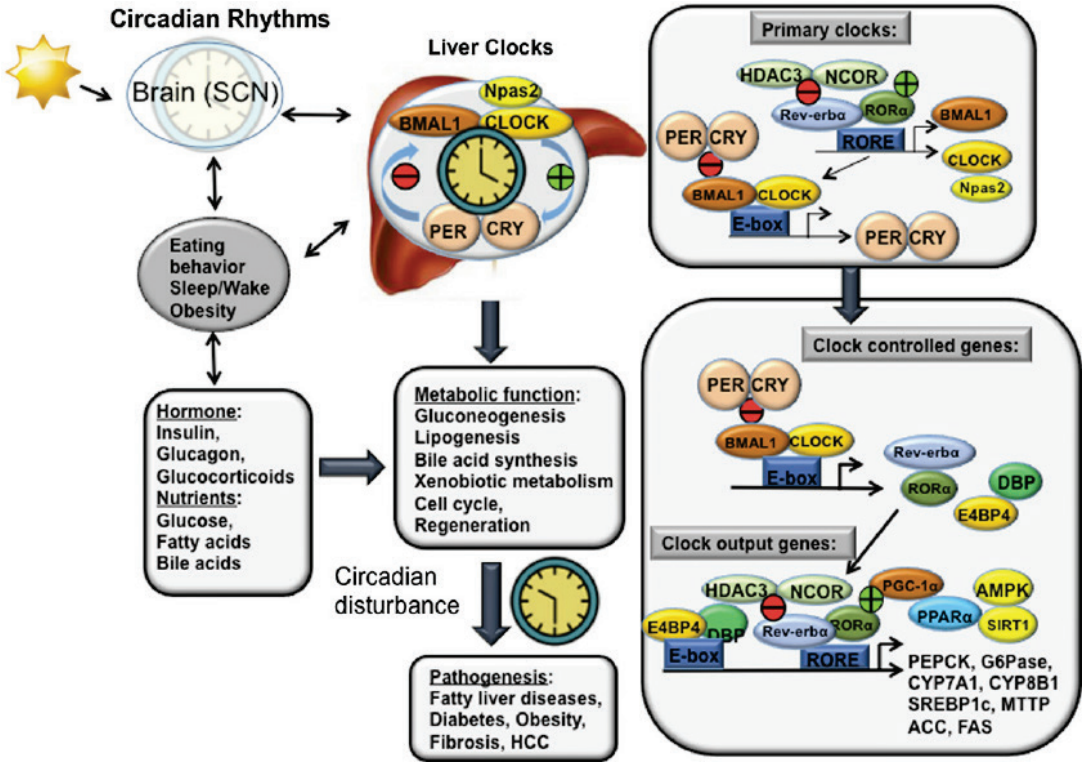
Circadian oscillations are entrained by several external factors, termed zeitgebers, such as the daily light-dark cycle[195], [196]. However, even when these are removed (e.g. experimentally), the oscillations continue for some time, which shows that the oscillations are regulated and maintained by an intrinsic clock of the body, rather than being caused directly by external factors[197].

A master clock, located in the supra-chiasmatic nucleus (SCN) of the brain, integrates the external factors and regulates the oscillations of peripheral tissues[198]. Peripheral oscillations can, however, also be influenced by other zeitgebers such as feeding and blood oxygen levels[199]–[201]. To refer to specific timepoints relative to the circadian clock, it is common to use either "zeitgeber time (ZT)" or "circadian time (CT)". ZT is defined relative to the start of the light phase (where ZT0 = light on and ZT12 = light off), while CT refers to the onset of activity of the organism (where CT0 = beginning activity in diurnal animals and CT12 = beginning of activity in nocturnal animals). In a light/dark cycle of 12/12 hours, the two timescales are congruent[202].

The liver, in particular, is known to be strongly influenced by circadian rhythms. A study comparing transcript oscillations in 12 mouse organs found that, out of all inspected organs, the liver had the most circadian transcripts[149]. The oscillating transcripts include not only components of the circadian clock but also critical metabolic pathways. Carbohydrate and lipid metabolism, as well as cholesterol and xenobiotic metabolism, are affected[203].

It should be mentioned that, although the circadian rhythm has a strong influence on liver gene expression, it is not the only factor determining temporal changes of gene expression. Feeding patterns and the metabolic state have especially a substantial effect on the liver as the main metabolic organ of the body[177], [204]. The intrinsic oscillations of peripheral tissues can be uncoupled from the master clock in the brain by restricted feeding patterns, and the phase of at least some processes can be completely reversed in mice by allowing food intake only during the light phase (when mice usually are inactive and have lower food intake)[199]–[201], [205]. Therefore, these factors are an additional cause for temporal

expression changes in the liver that are partially independent of the circadian clock. The liver dependence on circadian rhythms is visually explained in Background Figure 2.2 below.



Background Figure 2.2: Circadian rhythms in liver metabolism. The central clock in the SCN synchronises with the peripheral clock to regulate liver metabolisms. Eating behaviour, sleep/wake cycle, and obesity affect central clock and liver clock functions and their synchronisation. Hormones such as insulin, glucagon, and glucocorticoids, and nutrients including glucose, fatty acids, and bile acids affect circadian rhythms and liver metabolism. Bmal1 and Clock are primary clock products that bind to the E-box sequences in the Per and Cry gene promoters. Per and Cry complexes inhibit the Bmal/Clock complex in a negative loop to inhibit Per and Cry transcription. Bmal1 and Clock (also Npas2) are regulated by a negative regulator Rev-erb-α , and positive regulator ROR-α , which bind to the same ROR response element (RORE) in the promoters. Rev-erb a recruits HDAC3 and NcoR to inhibit gene transcription and ultimately the circadian rhythms of many CCGs, such as PEPCK and G6Pase in gluconeogenesis, CYP7A1 and CYP8B1 in bile acid synthesis, and SREBP-1c and MTTP in lipogenesis in the liver. Alteration in synchronisation of the central clock and liver clock contributes to the pathogenesis of fatty liver diseases, diabetes, and obesity, as well as fibrosis and hepatocellular carcinoma. HCC, hepatocellular carcinoma; MTTP, microsomal triglyceride transfer protein; NCOR, nuclear receptor corepressor; SCN, suprachiasmatic nucleus. Figure and caption are taken from *Li and Chiang, 2014*[206].

As many processes affected by circadian rhythms are also zonated, the question arises how spatial and temporal regulatory processes work together and influence each other. So far, there has not been a lot of work on circadian changes in liver zonation. However, there have been seemingly contradictory findings on the zonation of several metabolic processes, and it has been proposed in the past that some of these might be explainable by circadian influences on gene expression or protein regulation[207]–[210]. There are a couple of different scenarios of how spatial gene expression patterns could be

modulated by circadian rhythms. For instance, the expression of zonated genes could be regulated in an additive of multiplicative way, leading to different spatiotemporal patterns of expression. One could image that zonation changes in time, so that a gene might turn from periportal to pericentral, and reverse. In a more elaborate setting, one could also envision that circadian rhythms only influence the expression of a gene in a particular area of the liver lobule, causing sublobular oscillations of expression. The possibilities are almost infinite. In the study presented Section 2, part of the objective will be to describe which spatiotemporal scenarios of zonation are actually observed in the data.

1.3. Inference and model selection

1.3.1. Linear Mixed Models

1.3.1.1. Introduction

When dealing with longitudinal data, that is, data for which several measures are made on a given statistical unit, Linear Mixed Models (LMMs) can be used to correct for non-independence of the individual data points. LMMs are also called hierarchical models, as they allow to control for nested structure the data[211]. For instance, in the study presented Section 2, the data points representing mRNA expression in the different lobule layers are not always independent as they can originate from the same mouse.

In practice, one could deal with the linear dependence by merely aggregating the data, that is, taking the mean of all the data points for a given statistical unit (in our study, this would mean averaging mRNA expression over the different layers). However, often one is interested in the effect that stems in the longitudinal dimension (space in our study).

Another method to get rid of the linear dependence would be to run one regression per statistical unit. But this would mean dealing with unit-specific model and parameters, which is often something one wants to avoid. Also, since the individual regressions are made on fewer data points, the corresponding regressions are usually noisier.

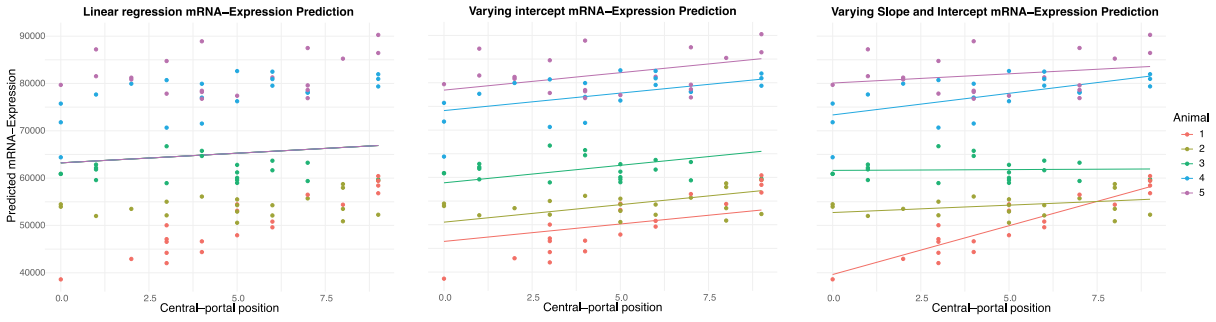
LMMs can be thought as a compromise between the two options, as LMMs has more parameters than a classic linear regression, but still incorporate the full structure of the data, enabling to reduce the noise of the individual estimates. LMMs have the other advantage of allowing easy comparison between the inter and intra differences among the statistical units.

1.3.1.2. Fixed and random effects

LMM have two types of parameters: those that explain how the data changes according to known sources of variability, also called fixed-effect parameters, and those that explain how the data changes from one statistical unit to another due to poorly understood (or uninteresting) factors, also called random-effects. As explicitly stated by their name, fixed-effects parameters stay the same from one statistical unit to another, while random-effect parameters do not.

In our study, rhythmicity and zonation parameters are all fixed-effects, as we suppose that they change with the genes, which happen to be identical across statistical units (the mice). Conversely, there's a lot of variability in gene expression among the animals. That could be due to the way the RNA-seq was done, or differences in how the animals sleep or eat, etc. There are many possibilities, but none of these is really of interest to us here, as we simply want to control for the inter-animal variability. Therefore, we added an intercept which was animal-specific in the model; this intercept is, by definition, a random-effect.

Background Figure 2.3 below shows three possibilities of modelling for a given dataset, in which several measures are made on different animals. Depending on the presence or absence of random-effects, the predictions made by the fixed-effects of the model can significantly vary, as well as the statistical confidence has in the corresponding parameters. Choosing the number of random effects can be a hard problem as it can only be driven by the preliminary knowledge of dependencies in the data.



Background Figure 2.3: Example of mixed-model regression on simulated hierarchical data. A total of 100 measures (mRNA-expression) is made on 5 animals, with 20 measures per animal. If the data is wrongly measured with a linear regression (left), therefore neglecting the dependency between the data points belonging to a given animal, one gets a non-significant slope ($pv = 0.7$) and a very significant intercept. If the intercept is assumed animal-specific (middle), the corresponding fixed-effect loses significance, but the slope becomes significant ($pv = 10^{-9}$). If one assumes that both the intercept and slopes are animal-specific (right), the slope is only barely significant ($pv = 0.02$). Simulations based on the code by Michael Freeman[212].

1.3.1.3. Theory

A mixed-model can be represented using a single equation using matrix notation:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad (i)$$

In this equation, the terms are as follow:

- $\boldsymbol{y}$ is the N-dimensional observation vector, whose mean is predicted by linear regression: $\overline{\boldsymbol{y}} = \boldsymbol{X\beta}$. N corresponds to the total number of observations, e.g. 20 if dealing with 5 data points per animal and 4 animals in totals.
- $\boldsymbol{\beta}$ is the p-dimensional vector of fixed-effects. p is usually reasonably low, e.g. 2 if considering 1 intercept, and 1 linear slope.
- $\boldsymbol{X}$ is the $N \times p$ dimensional design matrix of fixed effects, linking the observations $\boldsymbol{y}$ to the regression parameters of $\boldsymbol{\beta}$.
- $\boldsymbol{u}$ is the $(q \times J)$-dimensional vector of random-effects with mean 0 (and optimised variance-covariance matrix), where q is the number of parameters for the random effects and J the number of statistical units. q is usually low, as it's exceptional to consider more than 2 random parameters (e.g. random intercept and random slope). J is such that $\sum_j^J n_j = N$, where n_j correspond to the number of measures made on statistical unit j.
- $\boldsymbol{Z}$ is the $N \times (q \times J)$ design matrix of random effects, linking the observations $\boldsymbol{y}$ to the regression parameters of $\boldsymbol{u}$.
- $\boldsymbol{\varepsilon}$ is the N-dimensional error vector with mean 0 (and optimised variance)

In this equation, $\boldsymbol{\beta}, \boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ must be optimised on the data. In practice, $\boldsymbol{u}$ is not directly estimated, but assumed to be sampled from a multivariate Gaussian distribution with mean zero and unknown variance-covariance matrix $\boldsymbol{G}$:

$$\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{G}) \qquad (ii)$$

Therefore, it is actually $\boldsymbol{G}$ which must be optimised. As there is redundancy in the variance explained by $\boldsymbol{u}$ and $\boldsymbol{\beta}$, the optimisation requires non-linear methods, and since the matrix $\boldsymbol{Z}$ can be extremely large, the optimisation can be time-consuming.

As a side-note, model selection for mixed-model can be a hard problem as usual criteria such as BIC (see Section 4.4.2, in Chapter 1) are not well-defined for random-parameters. Thankfully, in recent years, several alternative methods have been developed, such as marginal AIC [213].

2. Published article

The article hereafter has been as little modified as possible. According to EPFL recommendations, the reference numbers are in the continuity of the thesis. Equation and Figure numbers are the same as in the published format (although the Chapter number is now indicated in the Figure captions).

2.1. Abstract

The mammalian liver performs key physiological functions for maintaining energy and metabolic homeostasis. Liver tissue is both spatially structured and temporally orchestrated. Hepatocytes operate in repeating anatomical units termed lobules and different lobule zones perform distinct functions. The liver is also subject to extensive temporal regulation, orchestrated by the interplay of the circadian clock, systemic signals and feeding rhythms. Liver zonation was previously analysed as a static phenomenon and liver chronobiology at the tissue level. Here, we use single-cell RNA-seq to investigate the interplay between gene regulation in space and time. Categorising mRNA expression profiles using mixed-effect models and smFISH validations, we find that many genes in the liver are both zonated and rhythmic, most of them showing multiplicative space-time effects. Such dually regulated genes cover key hepatic functions such as lipid, carbohydrate and amino acid metabolism. In particular, our data suggest that rhythmic and localised expression of Wnt targets may be explained by rhythmic Wnt signaling from endothelial cells near the central vein. Core circadian clock genes are expressed in a non-zonated manner, indicating that the liver clock is robust to zonation. Together, our comprehensive data reveal how liver function is compartmentalised spatio-temporally at the sub-lobular scale.

2.2. Introduction

The liver is a vital organ maintaining body physiology and energy homeostasis. The liver carries out a broad range of functions related to carbohydrate and lipid metabolism, detoxification, bile acid biosynthesis and transport, cholesterol processing, xenobiotics biotransformation, and carrier proteins secretion. Notably, the liver performs catabolic and anabolic processing of lipids and amino acids and produces the majority of plasma proteins[214]. Liver tissue is highly structured on the cellular scale, being heterogeneous in both cell-type composition and microenvironment[215]. In fact, liver tissue is made up of millions of repeating anatomical and functional subunits, called lobules, which in mice contain hepatocytes arranged in 12-15 concentric layers with a diameter of about 0.5mm[180]. On the portal side of the lobule, blood from the portal vein and the hepatic arteriole enters small capillaries called sinusoids and flows to the central vein. This is accompanied with gradients in oxygen

concentration, nutrients and signaling along the porto-central axis, with the latter notably involving the Wnt pathway[216], [217]. Due to this polarisation, hepatocytes in different layers perform separate functions, a phenomenon termed *liver zonation*[179], [214].

Recently, we combined single-cell RNA-sequencing (scRNA-seq) of dissociated hepatocytes and single-molecule RNA fluorescence in situ hybridisation (smFISH) to reconstruct spatial mRNA expression profiles along the porto-central axis[173]. This analysis revealed an unexpected breadth of spatial heterogeneity, with ~50% of genes showing spatially non-uniform patterns. Among them, functions related to ammonia clearance, carbohydrate catabolic and anabolic processes, xenobiotics detoxification, bile acid and cholesterol synthesis, fatty acid metabolism, targets of the Wnt and Ras pathways, and hypoxia-induced genes were strongly zonated.

In addition to its spatial heterogeneity, the liver is also highly dynamic temporally. Chronobiology studies showed that temporally gated physiological and metabolic programs in the liver result from the complex interplay between the endogenous circadian liver oscillator, rhythmic systemic signals, and feeding/fasting cycles[218], [219]·[220]. An intact circadian clock has repeatedly been demonstrated as key for healthy metabolism, also in humans[221]. Temporal compartmentalisation can prevent two opposite and incompatible processes from simultaneously occurring, for example, glucose is stored as glycogen following a meal and is later released into the blood circulation during fasting period to maintain homeostasis in plasma glucose levels. Functional genomics studies of the circadian liver were typically performed on bulk liver tissue[131]. In particular, we and others showed how both the circadian clock and the feeding fasting cycles pervasively drive rhythms of gene expression in bulk, impacting key sectors of liver physiology such as lipid and steroid metabolism[149], [222], [223].

Here, we asked how these spatial and temporal regulatory programs interact on the level of individual genes and liver functions more generally. In particular, can zonated gene expression patterns be temporally modulated on a 24 h time scale? And conversely, can rhythmic gene expression patterns observed in bulk samples exhibit sub-lobular structure? More complex situations may also be envisaged, such as time-dependent zonation patterns of mRNA expression (or, equivalently, zone-dependent rhythmic patterns), or sublobular oscillations that would escape detection on the bulk level due to cancelations. On the physiological level, it is of interest to establish how hepatic functions might be compartmentalised both in space and time. To study both the spatial and temporal axes, we performed scRNA-seq of hepatocytes at four different times along the 24 h day, extending our previous approach[173], [224] to reconstruct spatial profiles at each time point. The resulting space-time patterns were statistically classified using a mixed-effect model describing both spatial and temporal variations in mRNA levels. In total, ~5000 liver genes were classified based on their spatio-

temporal expression profiles, and a few representative profiles were further analysed with smFISH. Overall, this approach helped to elucidate the richness of space-time gene expression dynamics of the liver and provides a comprehensive view on how spatio-temporal compartmentalisation is utilised in the mammalian liver.

2.3. Results

2.3.1. Single-cell RNA-seq captures spatiotemporal gene expression patterns in mouse liver

To investigate spatio-temporal gene expression patterns in mouse liver, we sequenced mRNA from liver cells obtained from 10 mice at 4 different times of the day (ZT = 0h, 6h, 12h and 18h, two to three replicates per time point). We here focused on hepatocytes by enrichment of cells according to size and *in silico* filtering, yielding a total of 19663 cells (several filtering steps are involved, Methods). To validate that the expected axes of variation are present in the scRNA-seq data, we generated a low-dimensional representation of all cells (t-SNE projections) and colored cells either by their position along the centro-portal axis (layers) (Figure 2.1A) or time (Figure 2.1B). This revealed that known portally and centrally expressed transcripts, such as *Cyp2f2* (Figure 2.1C) or *Cyp2e1* (Figure 2.1D), respectively, mark cells in opposite regions of the projections. Likewise, time-of-day expression varied along an orthogonal direction, as shown for the *Elovl3* gene peaking at ZT0 (Figure 2.1E).

To reduce the complexity of the spatial variation in mRNA levels, we here introduced eight different lobule layers to describe gene expression along the centro-portal axis. For this, we adapted our previous method[224], with the modification that only transcripts that did not vary across time points were used as landmark zonated genes (Methods). The resulting reconstructed mRNA expression profiles yielded 80 (8 layers over 10 mice) data points for each transcript. These reconstructions faithfully captured reference zonated genes, with both central, and portal, expression (Figure 2.1F), such as *Glul,* and *Ass1*, respectively. The reconstructions also included reference core clock and rhythmic output genes, such as *Bmal1* (also named *Arntl*) and *Dbp*, with large differences in expression across the time points, and peaking at the expected times (Figure 2.1G). Finally, genes showing both zonated and rhythmic mRNA accumulation were found (Figure 2.1H), with both central (*Elovl3*) or portal (*Pck1*) expression patterns, and with specific peak times. Since most of the zonated profiles showed exponential shapes, and gene expression changes typically occur on a log scale[225], we log-transformed the data for further analysis (Methods, Figure 2.1F, Supplementary Figure 2.1A).

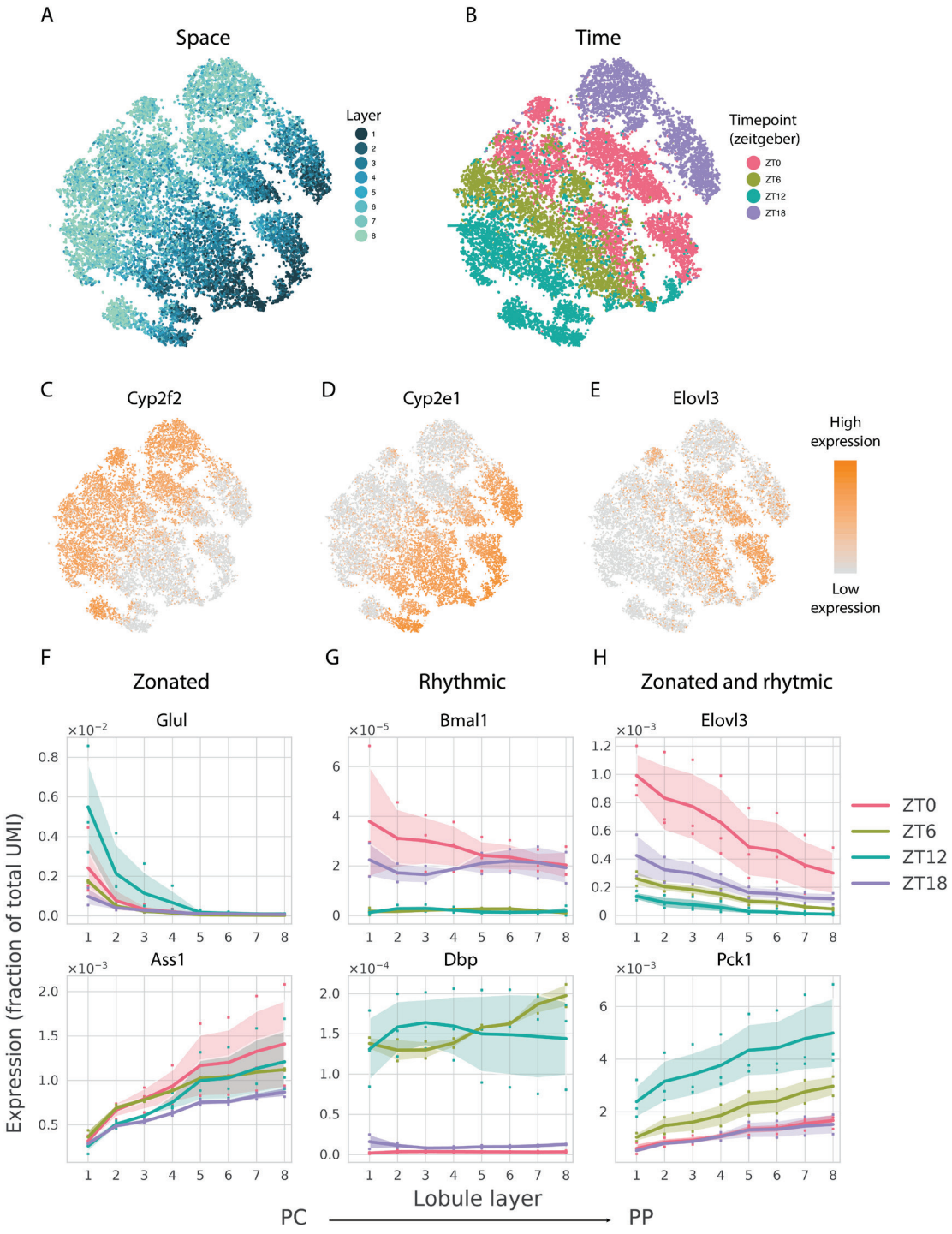


Figure 2.1: A scRNA-seq approach to space-time gene expression patterns in mouse liver. **(A-E)** *t*-SNE visualisations of the scRNA-seq data (19663 hepatocytes from n = 10 mice). Individual cells are colored by the lobule layer (**A**), Zeitgeber time (**B**), expression levels of the zonated genes *Cyp2f2* and *Cyp2e1* (**C-D**), or the circadianly regulated gene *Elovl3* (**E**). **(F-H)** Reconstructed spatial profiles (lobule layers 1-8) of selected zonated but temporally static genes (**F**, top: *Glul* pericentrally (PC) expressed, bottom: *Ass1* periportally (PP) expressed); rhythmic but non-zonated genes (**G**, top: *Bmal1* peaking at ZT0, bottom: *Dbp*, peaking at ZT6-12); zonated and rhythmic genes (**H**, Top: *Elovl3*, bottom: *Pck1*). Expression levels correspond to fraction of total UMI per cell in linear scale. Log-transformed profiles are in Figure 2.1. Dots in F-H represent data points from the individual mice. Shaded areas represent SD across mice.

2.3.2. Space-time mRNA expression profiles can be categorised according to zonation and rhythmicity

Next, we investigated if zonated gene expression patterns can be dynamic along the day, or conversely whether temporal expression patterns might be zone-dependent. To select a reliable set of reconstructed mRNA expression profiles for subsequent analyses, we first discarded lowly expressed genes, as well as genes with significant biological variability across replicate liver samples. This yielded 5058 spatio-temporal gene expression profiles (**Supplementary Figure 2.2**A). An exploratory analysis of variance clearly identified zonated genes, rhythmic genes, and fewer genes showing variability along both axes, with known zonated and rhythmic genes distributed as expected (Figure 2.2A).

To identify possible dependencies between spatial and temporal variations, we built a mixed-effect linear model[226] for the space-time mRNA profiles, which extends harmonic regression to include a spatial covariate (Figure 2.2B). In this model, rhythms are parameterised with cosine and sine functions, while spatial profiles are represented with (up to second order) polynomials. In its most complex form, the model uses nine parameters describing spatially modulated oscillations, and one intercept per mouse (Methods). When some of the parameters are zero, the model reduces to simpler mRNA profiles, for example purely spatial or purely temporal expression profiles (Figure 2.2C). We then used model selection[79] to identify the optimal parameterisation and category for each gene (Methods). *In fine*, we classified each mRNA profile into one of five types of patterns (Figure 2.2C). If only the intercept is used, the profile will be classified as flat (F). If only time-independent zonation parameters are retained, the predicted profile will be purely zonated (Z). If only layer-independent rhythmic parameters are retained, the predicted profile will be purely rhythmic (R). If only layer-independent rhythmic parameters and time-independent zonation parameters are retained, the profile is classified as independent rhythmic-zonated (Z+R). If at least one layer-dependent rhythmic parameter is selected, the profile will be termed interacting (ZxR). This classification revealed that, overall, about 30% of the mRNA profiles were zonated (Z, Z+R and ZxR) and about 20% were rhythmic (R, Z+R and ZxR) (Figure 2.2D). The peak times of these rhythmic transcripts were highly consistent with bulk chronobiology data[227] (Supplementary Figure 2.2B). This analysis is available as a web-app resource along with the corresponding data (https://czviz.epfl.ch).

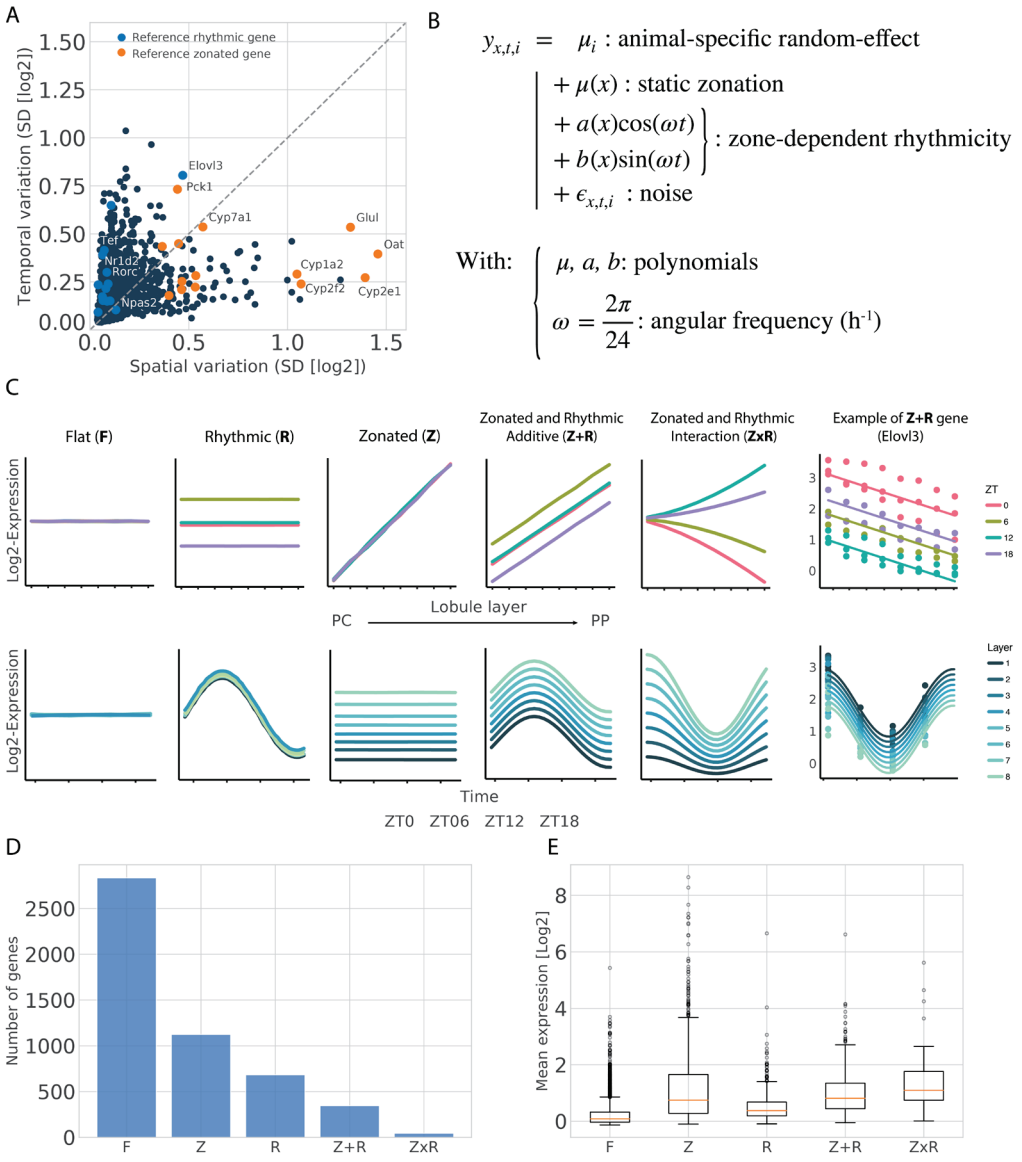


Figure 2.2: Space-time mRNA expression profiles categorised with mixed-effect models. (A) Spatial and temporal variation for each mRNA transcript profile, calculated as standard deviations (SD) of log2 expression along spatial or temporal dimensions (Methods). Colored dots correspond to reference zonated genes (orange) and reference rhythmic genes (blue) (Methods). **(B)** Generalised harmonic regression model for spatio-temporal expression profiles describing a static but zonated layer-dependent mean $\mu(x)$, as well as layer-dependent rhythmic amplitudes ($a(x)$ and $b(x)$). All layer-dependent coefficients are modeled as second order polynomials; i denotes the biological replicates. μ_i are random effects needed due to the asymmetric experimental design of the study (Methods). **(C)** Schema illustrating the different categories of profiles. Depending on which coefficients are non-zero (Methods), genes are assigned to: F (flat), Z (purely zonated), R (purely rhythmic), Z+R (additive zonation and rhythmicity), ZxR (interacting zonation and rhythmicity). Graphs emphasise either zonation (top), with the x-axis representing layers, or rhythmicity (bottom), with the x axis representing time (ZT). Right side of the panel: an example of fit (*Elovl3*). **(D)** Number of transcripts in each category. **(E)** Boxplot of the mean expression per category shows that the zonated genes (Z, Z+R and ZxR) are more expressed than flat (F) and rhythmic genes (R). Complex modulated genes (ZxR) are the most expressed according to median expression (orange line). Box limits are lower and upper quartile, whiskers extend up to the first datum greater/lower than the upper/lower quartile plus 1.5 times the interquartile range. Remaining points are outliers.

Interestingly, we found that 7% of the analysed genes in the liver were both zonated and rhythmic. Such dually regulated transcripts represent 25% of all zonated transcripts, and 36% of all rhythmic transcripts, respectively. The previously shown *Elovl3* transcript, involved in fatty acid elongation, and *Pck1*, a rate limiting enzyme in gluconeogenesis, are prototypical Z+R genes (Figure 2.1H). Gluconeogenesis requires acetyl-CoA produced via β-oxidation. As mice are in a metabolically fasted state towards the end of the light phase (~ ZT10) and oxygen needed for β -oxidation is most abundant portally[183] this process indeed needs to be both spatially and temporally regulated. Similarly, fatty acids production occurs during periods of excess energy and glycolysis (~ ZT18) and is located around the central vein[228]. The dual regulation of these genes may therefore ensure optimal liver function under switching metabolic conditions.

Dually regulated genes were mostly Z+R, with only a minority of ZxR patterns. The average expression across categories showed that rhythmic genes are more lowly expressed on average than genes in the other categories, likely reflecting shorter half-lives (Figure 2.2E and Supplementary Figure 2.2C). Together, we found that mRNA expression of many zonated genes in hepatocytes is not static, and is in fact compartmentalised both in space and time.

2.3.3. Properties of dually zonated and rhythmic mRNA profiles

The majority of dually regulated genes are Z+R, which denotes additive (in log) space-time effects, or dynamic patterns where slopes or shapes of spatial patterns do not change with time (Figure 2.2C). On the other hand, fully dynamic patterns (ZxR) are rare. Comparing the proportions of central, mid-lobular (peaking in the middle of the porto-central axis) and portal genes among the purely zonated genes (Z), and independently zonated and rhythmic genes (Z+R), did not reveal significant differences (Figure 2.3A), suggesting that rhythmicity is uncoupled with the direction of zonation. Similarly, comparing the phase distribution among the purely rhythmic genes (R) and the Z+R genes did not show a significant difference (Figure 2.3B), indicating that zonation does not bias peak expression time.

Moreover, oscillatory amplitudes were uncorrelated with the zonation slopes in Z+R genes (Figure 2.3C). However, we observed that large slopes (>1 or < -1, corresponding to fold changes of at least two between central and portal layers) are only associated with medium and large temporal amplitudes (>0.2, corresponding to >0.4 log2 fold change in time), as illustrated by *Elovl3*, *Rnase4* or *Pck1* (Supplementary Figure 2.3A). Conversely, many genes show small slopes and large amplitudes, as illustrated by *Hsp90aa1*, *Thrsp*, *Lpin1* (Supplementary Figure 2.3A).

Finally, for ZxR genes with potentially more complex space-time patterns, we investigated whether the spreads in amplitudes and peak times across the layers were linked (Figure 2.3D). This revealed that all ZxR profiles showed small spreads in peak times (<1h) but larger amplitude spreads (>0.1, log2). This phenomenon, illustrated by *Sds*, *Mt1* and *Cyp7a1* (**Supplementary Figure 2.3**B), indicates that amplitude modulation is the main factor for classification as ZxR.

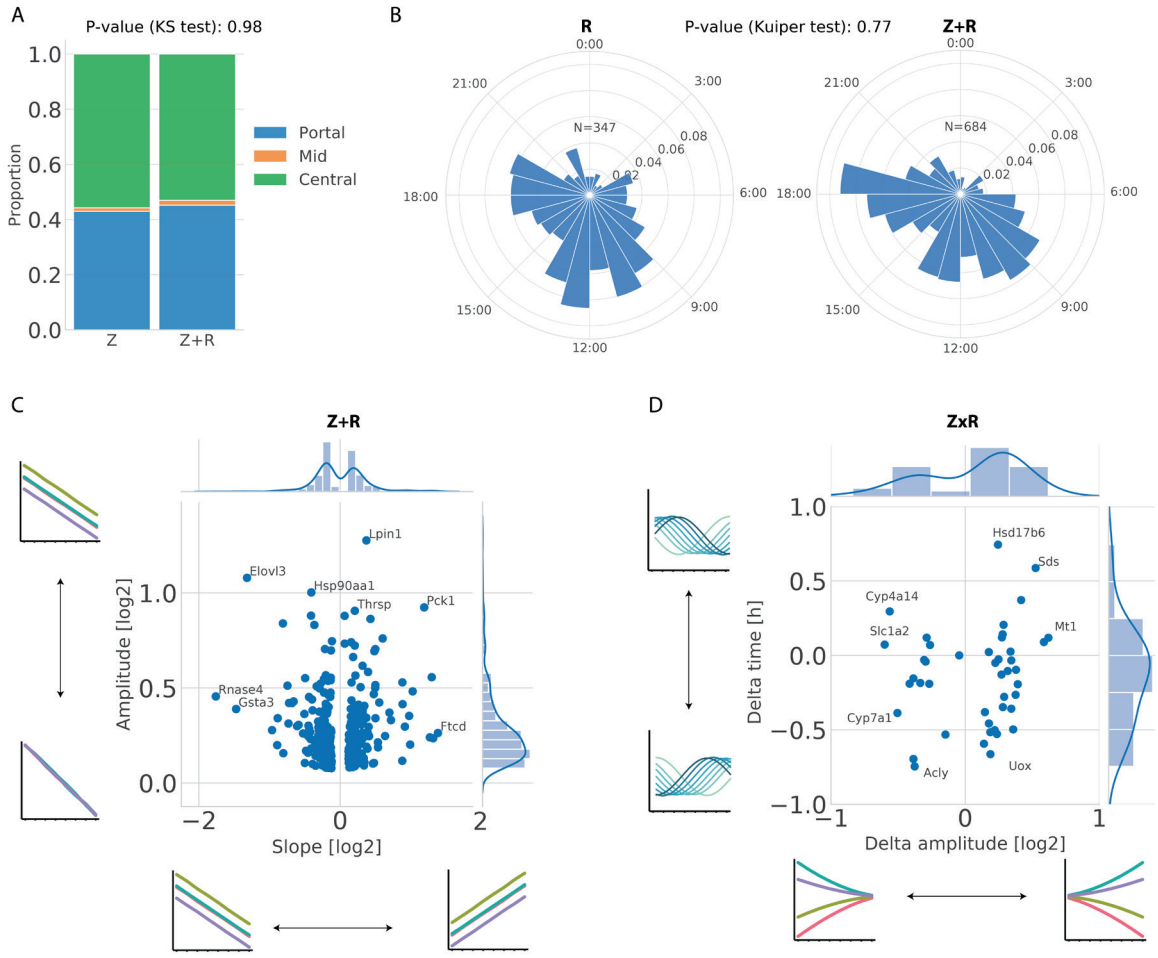


Figure 2.3: Properties of dually zonated and rhythmic mRNA profiles. (A) Proportions of pericentral (green) and periportal (blue) transcripts are similar in Z and Z+R. Mid-lobular genes (orange) are rare (<2%). **(B)** Peak time distribution of rhythmic transcripts are similar in R and Z+R categories. **(C)** Effect sizes of zonation (slope) *vs.* rhythmicity (amplitude) in Z+R genes. **(D)** Magnitude of time shifts (delta time, in hours) *vs.* amplitude gradient (delta amplitude, in log2) along the central-portal axis in ZxR genes.

2.3.4. smFISH analysis of space-time mRNA counts

To substantiate the RNA-seq profiles, we performed single RNA molecule fluorescence *in situ* hybridisation (smFISH) experiments on a set of selected candidate genes with diverse spatio-temporal patterns. smFISH provides a sensitive, albeit low-throughput measurement of mRNA expression. We selected the core-clock genes *Bmal1* and *Per1*, classified

as R in the RNA-seq analysis, *Elovl3* for Z+R, and *Acly* for the ZxR categories. Purely zonated genes (Z) were already well studied with smFISH[173].

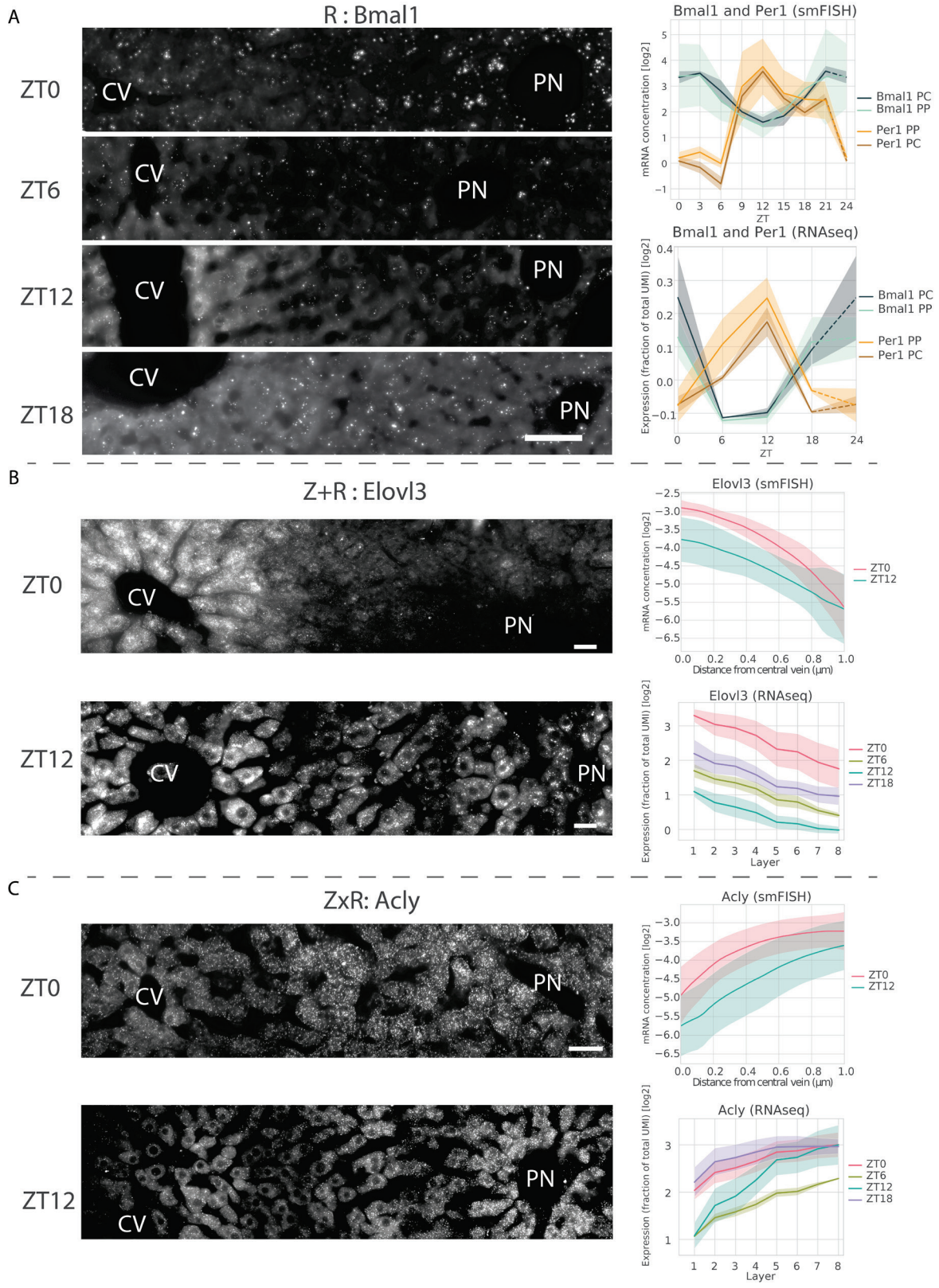


Figure 2.4: smFISH analysis of rhythmic and zonated transcripts. (A): smFISH (RNAscope, Methods) of the core clock genes *Bmal1* and *Per1* (both in R category) in liver slices sampled every 3

hours. **Left:** representative images at ZT0, ZT06, ZT12 and ZT18 for *Bmal1*. Pericentral veins (CV) and a periportal node (PN) are marked. Scale bas is 50µm. Endothelial cells lining the PC and cholangiocytes surrounding the PP were excluded from the quantification. mRNA transcripts and nuclei were detected in PN and PC zones (Methods). **Right:** temporal profiles of *Bmal1I* and *Per1* from smFISH (top, quantification of the number of mRNA transcripts at 8 time points ZT0 to ZT21, every 3 hours, shaded area indicate SD across images), in PN and PC regions, and scRNA-seq (bottom, shaded areas is SD across mice). **(B-C)** smFISH (Stellaris, Methods) for *Elovl3* (Z+R) and *Acly* (ZxR). smFISH quantifications were made for ZT0 and ZT12 (Methods). **Left:** representative images at ZT0, ZT12 for *Elovl3* (B) or *Acly* (C). Pericentral veins (CV) and a periportal node (PN) are marked. Scale bar - 20µm. **Right:** quantified profiles for each gene in the two time points from smFISH (top, shaded area indicate SD across images), prediction of profiles at four time points base on the scRNA-seq (bottom, shaded areas is SD across mice).

The reconstructed scRNA-seq and smFISH profiles were consistent, though with minor discrepancies. *Bmal1* (~ZT0) and *Per1* (~ZT12) phases were nearly identical in both experiments, and the rhythms did not depend on the lobular position (Figure 2.4A). *Elovl3* is both centrally biased and rhythmic in RNA-seq and smFISH, even though the amplitude of the oscillations appeared lost on the portal side in the FISH experiment, presumably due to low expression and thus low signal-to-noise (Figure 2.4B). Finally, *Acly* showed a pattern in smFISH data which validates its classification as ZxR, especially since the amplitude is lower on the portal side where the transcript is more highly expressed (Figure 2.4C).

2.3.5. Space-time logic of hepatic functions and activity of signaling pathways

We next used our classification to understand the spatio-temporal dynamics of hepatic functions and signaling pathways in the liver. Given the prevalence of zonated gene expression profiles, we first analysed if the circadian clock is sensitive to zonation. We found that profiles of reference core-clock genes (Supplementary Figure 2.4) were assigned to the rhythmic only category (R), except for *Cry1* and *Clock* that were assigned to Z+R, but with high probabilities also for R (Supplementary Tables 2.1 and 2.4). This suggests that the circadian clock is largely non-zonated and therefore robust to the heterogenous hepatic microenvironment.

We then systematically explored enrichment of biological functions in the R, Z, and Z+R categories using DAVID[229] (Supplementary Table 2.2). Gene clusters related to the circadian clock were clearly enriched in the R category, however, no other functions stood out as purely rhythmic. Functions of zonated genes have been described previously [173], [230], here, our analysis of Z only genes highlighted that processes related to protein secretion were highly enriched in portal genes, constituents of ribosomes were strongly biased centrally, as were many P450 enzymes involved in oxidation of steroids, fatty acids and xenobiotics. Among the dually regulated Z+R genes, lipid metabolism stood out as the most enriched

function, with about two thirds of central and one third of portal gene expression, showing peak times that were mostly between ZT9 and ZT21 (Figure 2.5A). Moreover, peroxisome related genes peaked in a similar interval with a marked preference for central patterns (Figure 2.5B). Lastly, genes linked with protein glycosylation peaked throughout the day and showed more portal patterns, presumably linked to portally biased protein secretion (Figure 2.5C). A similar analysis of enrichment in KEGG pathways, using a more exhaustive collection of backgrounds sets (Supplementary Table 2.3) confirmed that rhythmic gene expression is spatially flat for core clock function, but zonated (Z+R) for other hepatic functions.

Next, we investigated genes targeted by reference signaling pathways previously implicated in zonation[173]. Specifically, we considered genes targeted by the Wnt, Ras and hypoxia pathways. As expected, the three pathways where strongly enriched in zonated genes (Supplementary Figure 2.5, column B5). In agreement with ref[173], both the positive targets of Wnt and the negative targets of Ras are strongly enriched in central genes, while the negative and positive targets, respectively, are strongly enriched in portal genes (Supplementary Figure 2.5, column B4). However, no such bias was found for hypoxia. Regarding day vs. night expression, the Wnt and hypoxia targets displayed strong bias, while Ras did not (Supplementary Figure 2.5, column B1). Moreover, Wnt targets (both positive and negative) as well as positive targets of hypoxia showed rhythms preferentially peaking during the day (Supplementary Figure 2.5, column B2). Among the Z+R genes, the enrichment patterns of central (noted Z^C+R) and portal (Z^P+R) genes were similar during night and day (Supplementary Figure 2.5, column B3). On a finer temporal scale (Figure 2.6A), the rhythmic targets (R and Z+R) of Wnt and hypoxia showed a common pattern of temporal compartmentalisation: the negative targets tended to be enriched around ZT0 (dark-light transition) and underrepresented around ZT14, while the positive targets were enriched around ZT10 and underrepresented around ZT20 (until about ZT3 for Hypoxia). Ras targets, positive or negative, did not exhibit significant temporal bias.

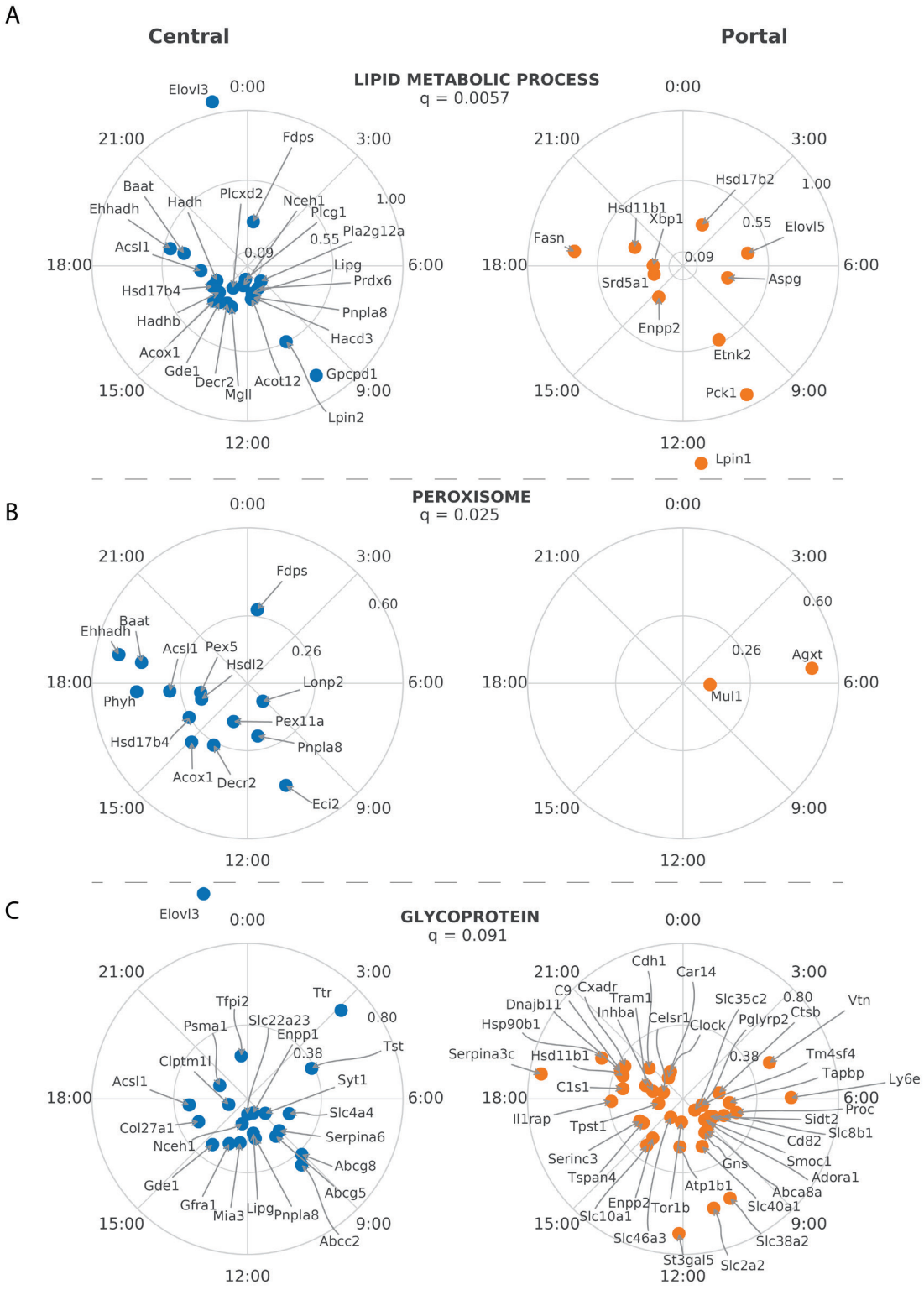


Figure 2.5: Space-time logic of compartmentalised hepatic functions for Z+R genes. (A-C) Polar representations of central (**left**) and portal (**right**) genes from three prominently enriched pathways in the Z+R category (Supplementary Table 2.2). Peak expression times are arranged clockwise (ZT0 on vertical position) and amplitudes (log2, values indicated on the radial axes) increase radially. q indicates significance of enrichment (Bonferroni adjusted).

Finally, we asked whether the temporal oscillations in the expression of Wnt-activated genes might correlate with temporal oscillations in the Wnt morphogens produced by pericentral liver endothelial cells (LECs). To this end, we performed smFISH experiments and quantified the expression of the Wnt ligand *Wnt2*, as well as the Wnt antagonist *Dkk3* (Figure 2.6B). Consistent with the enrichment of peak times in positive Wnt targets (Figure 2.6A), we found that *Wnt2* expression in LECs exhibited non-uniform expression around the clock, with a significant trough at ZT18 (p=0.0007, Kruskal-Wallis). Although differences in *Dkk3* expression were not significant (p=0.3), the observed median expression was lowest at ZT12, when the expression of pericentral Wnt-targets is the highest.

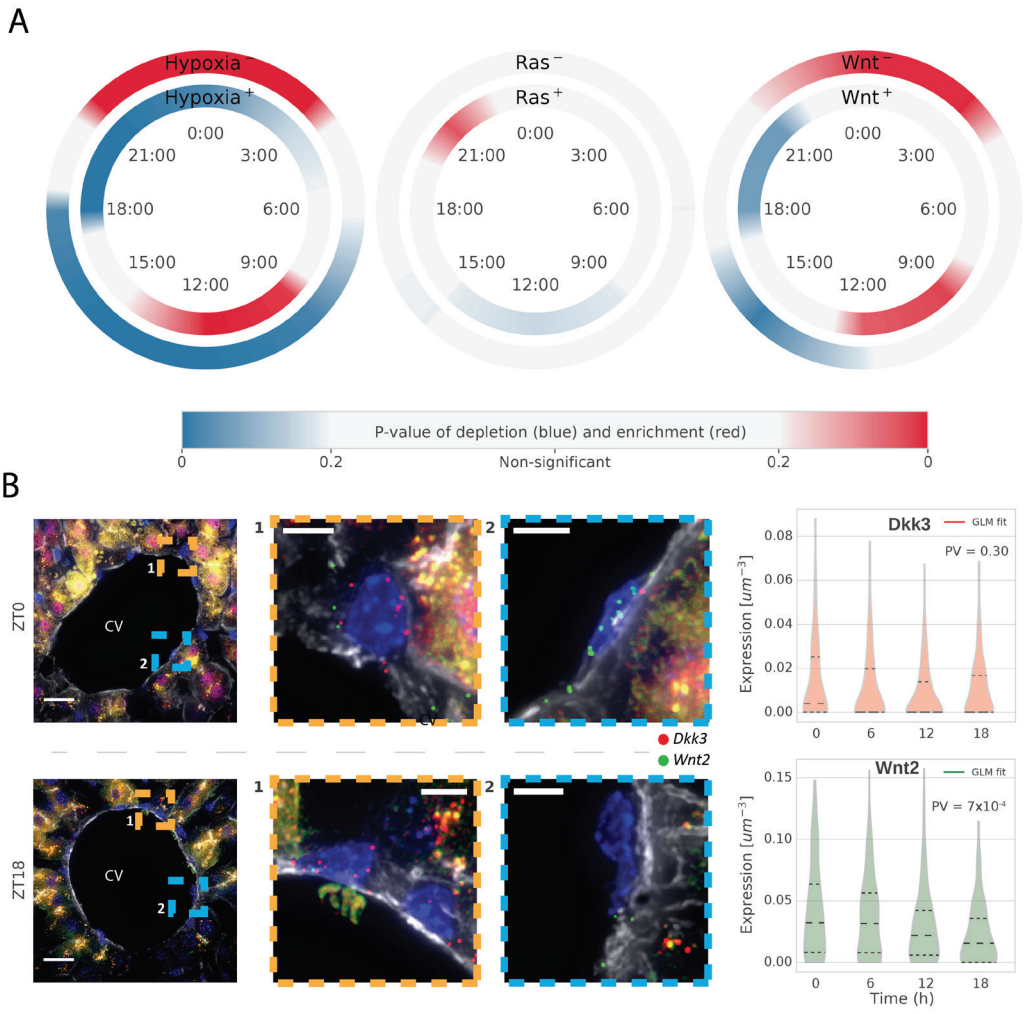


Figure 2.6: Wnt-targets correlates with Wnt2 expression. A) Enrichment/depletion of genes targeted by the Wnt, Ras and Hypoxia pathways with respect to the rhythmic genes (R and Z+R) peaking at different time of the day (window size: 3h). Colormap show p-values (two-tailed hypergeometric test): red (blue) areas indicate times of the day for which there are more (less) pathways genes peaking than expected. **(B) Left:** Representative smFISH images of *Wnt2* and *Dkk3* expression in endothelial cells at ZT0 (top) and ZT18 (bottom), nuclei are stained in blue (DAPI) and cell membrane in grey (phalloidin). Scale bars, 20 µm; inset scale bars 5 µm; CV = central vein. **Right:** Quantitative analysis of smFISH images (516 cells of 120 central veins of at least two mice per time point). mRNA expression is in smFISH dots per µm^3. Dashed lines correspond to quartiles.

2.4. Discussion

Recent genome-wide analyses of zonated gene expression in mouse and human liver[173], [231] uncovered a rich organisation of liver functions in space at the sublobular scale, while chronobiology studies of bulk liver tissue revealed a complex landscape of rhythmic regulatory layers orchestrated by a circadian clock interacting with feeding-fasting cycles and systemic signals[223], [232]–[234]. Here, we established how these two regulatory programs intertwine to shape the daily space-time dynamics of gene expression patterns and physiology in adult liver by extending our previous scRNA-seq approach[173]. We found that liver function uses gene expression programs where many genes exhibit compartmentalisation in both space and time.

We chose to focus on the parenchymal cells in the liver, the hepatocytes, for which smFISH data on landmark zonated genes was readily available, which enabled reconstructing spatio-temporal mRNA profiles from scRNA-seq[173]. Our approach may be extended to other cell types in the liver; in fact, static zonation mRNA expression profiles have been obtained for LEC, using a paired-cell approach[224]. In addition, *ab initio* reconstruction methods such as diffusion pseudo time[231] or novoSpaRc[235] could be used for spatially sparse cell types with no available zonated marker genes, e.g. stellate or resident immune Kupffer cells.

To study whether the observed space-time expression profiles may be regulated by either liver zonation, 24h rhythms in liver physiology, or both, we developed a mixed-effect model, combined with model selection. This enabled classifying gene profiles into five categories representing different modes of spatio-temporal regulation, from flat to wave-like. To validate these, we performed smFISH in intact liver tissue, which showed largely compatible profiles although some quantitative differences were observed. These differences most likely reflect the limited sensitivity of RNA-seq, uncertainties in the spatial analysis of smFISH in tissues, as well as known inter-animal variability in the physiologic states of individual livers, notably related to the animal-specific feeding patterns[227].

Together, this temporal analysis confirms that a large proportion of gene expression in hepatocytes is zonated[173] or rhythmic[204], and in addition reveals marked spatio-temporal regulation of mRNA levels in mouse liver (Z+R and ZxR genes, comprising 7% of all detected genes according to our criteria). This means that zonated gene expression patterns can be temporally modulated on a circadian scale, or equivalently, that rhythmic gene expression patterns can exhibit sub-lobular structure. The dominant mode for dually regulated gene was Z+R, which corresponds to additive effects of space and time in log, or multiplicate effects in the natural coordinates, and describes genes expression patterns that are compartmentalised in both space and time. In other words, such patterns are characterised by shapes

(in space) that remain invariant with time, but whose amplitudes are rhythmically rescaled in time. Or equivalently, the oscillatory amplitude (fold change) and phases are constant along the lobular coordinate, but the mean expression is patterned along the lobule. Such multiplicative effects could reflect the combined actions of transcriptional regulators for the zone and rhythm on promoters and enhancers of Z+R genes. Indeed, gene expression changes induced by several regulators combine multiplicatively[225]. The non-zonated expression of the core clock genes we identified is compatible with such a uniform non-zonated multiplicative effect. Note that though the (relative) shape of Z+R patterns is invariant in time, threshold-dependent responses that would lie downstream of such genes would then acquire domain boundaries which can shift in time. Finally, space-time waves of gene expression (ZxR) were also observed, but to a much lesser extent, and usually with larger amplitude than phase modulation along the lobular layers.

In addition to previously discussed zonated functions[173], pathway analysis revealed that expression of ribosome protein genes is higher centrally, which, together with the previously noted zonation of proteasome components[173] could indicate that protein turnover is higher in the pericentral lobule layers. This higher turnover could preserve protein function in the stressed low-oxygen and xenobiotics-rich pericentral microenvironment. *Fatty acid metabolism* appears complex: Z+R genes involved in fatty acid oxidation/degradation (*Acl1*, *Acox1*, *Ehhadh*, *Hacd3*, *Hadh*, *Hadhb*) are expressed centrally and peak around ZT12-ZT18, while *fatty acid elongation* genes either central (*Elovl3*) or portal (*Fasn*, *Elovl5*), peaking at different times of day. Interestingly, we found that expression of circadian clock transcripts is robust to metabolic zonation and is the only function showing an over-representation of rhythmic but non-zonated genes.

It was previously shown that Wnt signaling can explain the zonation of up to a third of the zonated mRNAs[7]. Wnt ligands are secreted by pericentral LECs [216], [217, p. 43] and form a graded spatial morphogenetic field. As a result, and as observed in our enrichment analysis (Supplementary Figure 2.5), Wnt-activated genes were pericentrally-zonated, whereas Wnt-repressed genes periportally-zonated. Our smFISH analysis suggested that temporal fluctuations in the expression of key ligands by pericentral LECs might account for oscillatory and zonated hepatocyte gene expression.

In summary, we demonstrate how liver gene expression can be quantitatively investigated with spatial and temporal resolution and how liver function is compartmentalised along these two axes. In particular our data suggest two scenarios: multiplicative effects of spatial and temporal regulators, and temporal regulation of spatial regulators such as WNTs.

2.5. Material and methods

2.5.1. Animals and ethics statement

All animal care and handling were approved by the Institutional Animal Care and Use Committee of WIS and by the Canton de Vaud laws for animal protection (authorisation VD3197.b). Male C57bl6 mice aged of 6 weeks, housed under reverse-phase cycle and under ad libitum feeding were used to generate sc-RNA-seq data of hepatocytes and single-molecule RNA-FISH (smFISH). Male mice between 8 to 10 weeks old, housed under 12:12 light-dark cycle, and having access to food only during the night (restricted-feeding) were used for smFISH of circadian clock genes.

2.5.2. Hepatocytes isolation and single-cell RNA-seq

Liver cells were isolated using a modified version of the two-step collagenase perfusion method of Seglen[236]. The tissue was digested with Liberase Blendzyme 3 recombinant collagenase (Roche Diagnostics) according to the manufacturer instructions. To enrich for hepatocytes, we applied a centrifuge step at 30g for 3 min to pull down all hepatocytes while discarding most of the non-parenchymal cells that remained in the sup. We next enriched for live hepatocytes by percoll gradient, hepatocytes pellet was resuspended in 25 ml of PBS, percoll was added for a final concentration of 45% and mixed with the hepatocytes. Dead cells were discarded after a centrifuge step (70g for 10min) cells were resuspended in 10x cells buffer (1x PBS, 0.04% BSA) and proceeded directly to the 10x pipeline. The cDNA library was prepared with the 10X genomics Chromium system according to manufactures instructions and sequencing was done with Illumina Nextseq 500.

2.5.3. Filtering of raw scRNA-seq data

The initial data analysis was done in R v3.4.2 using Seurat v2.1.0[237]. Each expression matrix was filtered separately to remove dead, dying and low quality cells. We firstly only kept genes that were expressed in at least 5 cells in any of the ten samples. We then defined a set of valid cells with more than 500 expressed genes and between 1000 and 10000 unique molecular identifiers (UMIs) and secondly an additional expression matrix with cells having between 100 and 300 UMIs which was used for background estimation. Other UMI-filters have been tried, but yielded equal or less reliable profiles. The mean expression of each gene was then calculated for the background dataset and subtracted from the set of valid cells. This was subsequently filtered to only include hepatocytes by removing cells with expression of non-parenchymal liver cell genes. Next, the cells were filtered based on the fraction of mitochondrial gene expression. First, expression levels in each cell were normalised

by the sum of non-mitochondrial and non-major urinary protein (*Mup*) genes. Indeed, as mitochondria are more abundant in periportal hepatocytes, the expression of mitochondrial genes is higher in this area[238]; and since these genes are very highly expressed, including them would reduce the relative expression of all other genes based on the cell's lobular location. *Mup* genes are also highly abundant and mapping their reads to a reference sequence is unreliable due to their high sequence homology[239]. Then, to assess the quality of the selected sets of cells, we computed the anti-correlation between the expression levels of *Cyp2e1* and *Cyp2f2*, two strongly and oppositely zonated genes, across all cells. Cells with 9-35% mitochondrial gene expression yielded the best quality, and we used these as input for the spatial reconstruction algorithm.

2.5.4. Spatial reconstruction of zonation profiles from scRNA-seq data

2.5.4.1. Choice of landmark genes.

The reconstruction algorithm relies on *a priori* knowledge about the zonation of a small set of landmark genes to infer the location of the cells. Reference [173] used smFISH to determine the zonation pattern *in situ* of 6 such landmark genes and used them to reconstruct the spatial profiles of all other genes at a single time point. Since we here aimed at reconstructing zonation profiles at different time points, we could not rely on those landmark genes, which might be subject to temporal regulation. Therefore, we used an alternative strategy where we selected landmark zonated genes from Reference [173] ($q < 0.2$), with the additional constraints that those should be highly expressed (mean expression in fraction UMI of more than 0.01% and less than 0.1%), and importantly vary little across mice and time. Specifically, we calculated the variability in the mean expression (across all layers) between all mice for every gene and removed genes with >= 10% variability. This yielded 27 central (*Akr1c6, Alad, Blvrb, C6, Car3, Ccdc107, Cml2, Cyp2c68, Cyp2d9, Cyp3a11, Entpd5, Fmo1, Gsta3, Gstm1, Gstm6, Gstt1, Hpd, Hsd17b10, Inmt, Iqgap2, Mgst1, Nrn1, Pex11a, Pon1, Psmd4, Slc22a1, Tex264*); and 28 portal (*Afm, Aldh1l1, Asl, Ass1, Atp5a1, Atp5g1, C8a, C8b, Ces3b, Cyp2f2, Elovl2, Fads1, Fbp1, Ftcd, Gm2a, Hpx, Hsd17b13, Ifitm3, Igf1, Igfals, Khk, Mug2, Pygl, Sepp1, Serpina1c, Serpina1e, Serpind1, Vtn*) landmark genes.

2.5.4.2. Reconstruction algorithm.

The reconstruction algorithm is based on the algorithm in reference [173] and was used in the modified version from reference [224]. The procedure was applied independently on each mouse, yielding ten spatial gene expression profiles for each gene, given as fraction of UMI per cell.

2.5.5. Spatiotemporal analysis of liver gene expression profiles

2.5.5.1. Data

Each profile for the 14678 genes includes 8 layers from the pericentral to the periportal zone and 4 time points: ZT0 (n=3 biological replicates from individual mice), ZT6 (n=2), ZT12 (n=3) and ZT18 (n=2). The expression levels (noted as x) are then log-transformed as follows:

$$y = log_2(x + \Delta) - B \quad (i)$$

The offset $\Delta = 10^{-4}$ buffers variability in lowly expressed genes, while the shift $B = -log_2(11 \times 10^{-5})$ changes the scale so that $y = 0$ corresponds to about 10 mRNA copies per cell (we expect on the order of 1M mRNA transcripts per liver cell).

2.5.5.2. Reference genes

For ease of interpretation (Figure 2.2 and **Supplementary Figure 2.2**), we used a set of reference circadian genes and a set of reference zonated genes, highlighted in several figures.

The reference core circadian clock and clock output genes are the following: *Bmal1, Clock, Npas2, Nr1d1, Nr1d2, Per1, Per2, Cry1, Cry2, Dbp, Tef, Hlf, Elovl3, Rora, Rorc.*

The reference zonated genes are the following: *Glul, Ass1, Asl, Cyp2f2, Cyp1a2, Pck1, Cyp2e1, Cdh2, Cdh1, Cyp7a1, Acly, Alb, Oat, Aldob, Cps1.*

2.5.5.3. Gene expression variance in space and time

To analyse variability in space and time (Figure 2.2A) we computed, for each gene, the spatial variance V_x and the temporal variance V_T. Let $y_{x,t,j}$ represent the expression profile, with j the replicate index, $t \in \{1,2,\dots,N_t\}$ the time index, and $x \in \{1,2,\dots,N_x\}$ the layer index. Then, V_x and V_T are computed as follows:

$$V_X = \frac{1}{N_t}\sum_t \frac{\sum_x [\sum_j (y_{x,t,j} - \frac{1}{N_x}\sum_x y_{x,t,j})]^2}{N_j^2 N_x} \quad (ii)$$

$$V_T = \frac{1}{N_x}\sum_x \frac{\sum_t [\sum_j (y_{x,t,j} - \frac{1}{N_t}\sum_t y_{x,t,j})]^2}{N_j^2 N_t} \quad (iii)$$

Thus, the spatial variance V_x is computed along the space (and averaged over the replicates) for each time condition, and then averaged over time. The procedure is similar, symmetrically, for V_t.

2.5.5.4. Genes filtering

For the analyses in Figure 2.2, we selected transcripts that were reproducible between replicates, as well as sufficiently highly expressed (see scatterplot in **Supplementary Figure 2.2**A). To assess reproducibility across replicates, we computed the average relative variance of the spatiotemporal profiles over the replicates:

$$V_J = \frac{\frac{1}{N_x N_t}\sum_{x,t}\left[\frac{1}{N_j}\sum_j\left(y_{x,t,j} - \frac{1}{N_j}\sum_j y_{x,t,j}\right)^2\right]}{\frac{1}{N_x N_t N_j}\sum_{x,t,j}\left[(y_{x,t,j} - \frac{1}{N_x N_t N_j}\sum_{x,t,j} y_{x,t,j})^2\right]} \quad (iv)$$

We considered genes with values below 50% (**Supplementary Figure 2.2**). To filter lowly expressed genes, we required the maximum expression level across layers and time points (fraction of UMIs) to exceed 10^{-5} which corresponds to y = 0 or about 10 copies of mRNA per cell. While this kept most of the reference zonated and circadian genes, the filtering selected a total of 5085 genes, used for all analyses presented in Figures 2-5. In addition, we systematically discarded *Cyp2a4* and *Cyp2a5* from analyses, as these genes are not discernable from scRNA-seq due to their highly similar sequences.

2.5.5.5. Mixed-effect model for spatiotemporal mRNA profiles

Since the data is longitudinal is space (8 layers measured in each animal), modelling the space-time profiles require the use of mixed-effect models. To systematically analyse the spatiotemporal mRNA profiles, we used a parameterised function. Specifically, the model uses sine and cosine functions for the time, and polynomials (up to degree 2) for space. Possible interaction between space and time are described as space-dependent oscillatory functions, or equivalently, time-dependent polynomial parameters. Our model for the transformed mRNA expression y reads:

$$y_{x,t,i} = \mu_i + \mu(x) + a(x)\cos(\omega t) + b(x)\sin(\omega t) + \varepsilon_{x,t,\mathrm{i}} \quad (v)$$

Here t is the time, x the spatial position along the liver layers, and $i \in \{1,2,\ldots,10\}$ the animal index. This function naturally generalises harmonic regression, often used for analysis of circadian gene expression[227], by introducing space-dependent coefficients:

$$\begin{cases} \mu(x) &= \mu_0 + \mu_1 P_1(x) + \mu_2 P_2(x) \\ a(x) &= a_0 + a_1 P_1(x) + a_2 P_2(x) \\ b(x) &= b_0 + b_1 P_1(x) + b_2 P_2(x) \end{cases}$$

Here, P_1 and P_2 are the Legendre polynomials of degrees 1 and 2, respectively; μ_0, μ_1 and μ_2 represent the static zonation profile, a_0 and b_0 represent the global (space-independent) rhythmicity of the gene, while a_1, a_2, b_1, b_2 represent layer-dependent rhythmicity. $\varepsilon_{x,t,j}$ is a Gaussian noise term with standard deviation σ. In addition to the fixed-effect parameters described so far, we also introduced a mouse-specific random-effect μ_i (with zero mean). This parameter groups the dependent layer measurements (obtained in the same animal) and thereby properly adjusts the biological sample size for the rhythmicity analysis.

Phases φ (related peak times t through $t = \varphi * 24/2\pi$) and amplitudes A, for each profile can then be computed for any layer from the coefficients $a(x)$ and $b(x)$:

$$\varphi(x) = arctan2\big(b(x), a(x)\big) \qquad A(x) = \sqrt{a(x)^2 + b(x)^2} \qquad (vi)$$

We also note that an equivalent writing of the model formulates the problem in terms of time-dependent zonation parameters instead of space-dependent rhythmicity:

$$y_{x,t,i} = \mu_i + \mu_0(t) + \mu_1(t)P_1(x) + \mu_2(t)P_2(x) + \varepsilon_{x,t,i} \qquad (vii)$$

where:

$$\begin{cases} \mu_0(t) &= \mu_0 + a_0\cos(\omega t) + b_0\sin(\omega t) \\ \mu_1(t) &= \mu_1 + a_1\cos(\omega t) + b_1\sin(\omega t) \\ \mu_2(t) &= \mu_2 + a_2\cos(\omega t) + b_2\sin(\omega t) \end{cases}$$

In this study, we fixed $\omega = \frac{2\pi}{24\ h}$ since the animals were entrained in a 24 h light-dark cycle and the low time resolution would prevent us from studying ultradian rhythms.

The model parameters, including the variance of the random effects and Gaussian noise strength σ, are estimated for each gene using the *fit* function from the Python library Stats-Models (version 0.9.0). Nelder-Mead was chosen as the optimisation method, and the use of a standard likelihood was favored over the REML likelihood to allow for model comparison[240]. To prevent overfitting of the gene profiles, we added a noise offset $\sigma_0 = 0.15$ [log_2] to the estimated noise σ, in the expression of the likelihood function used in the mixed-effect model optimisation.

Depending on the gene, the model presented in (v) and (vii) may be simplified by setting all or some of the (fixed) parameters to 0. For example, a non-oscillatory gene profile would normally have non-significant a_j and b_j parameters. In practice, considering the fixed effects, 2^9 sub-models of various complexity can be generated. However, we added a few reasonable requirements to reduce the number of models. First, the intercept μ_0 must be present in every model. Similarly, the parameters a_0 and b_0, providing a global rhythm, must be present

in every rhythmic model. Finally, the parameters a_j and b_j for j=0,1,2 must be paired to ensure a proper phase definition (*vi*).

The models can then be classified in different categories, depending on the retained (non-zero) parameters (Figure 2.2):

- The model comprising only the intercepts μ_0 and μ_i, termed flat (F).
- The models comprising only the intercepts and zonation parameters: $\mu_1$and/or μ_2, termed purely zonated (Z).
- The models comprising only the intercepts and rhythmic parameters: a_0 and b_0, termed purely rhythmic (R).
- The models comprising only the intercepts, zonated parameters and rhythmic parameters: μ_1 and/or μ_2, and a_0, b_0, termed independent (Z+R).
- The models comprising interaction parameters: $a_j$and b_j for j=1,2, termed interacting (ZxR).

Note that, since the random-effect parameter μ_i is interpreted as a correction for the data rather than a predictive parameter, it is systematically used for the fits, but not plotted in the final retained profiles (e.g. Figure 2.1, Supplementary Figure 2.1).

The Bayesian Information Criterion (BIC) is then used for model selection, enabling to choose the best model for a given gene profile. It appears that, for some profiles, several competing models can result in very close BIC values (see e.g. the discussion on *Clock* and *Cry*1 in the Results). Therefore, if some models have a relative difference of less than 1% in their BIC (sorted by increasing order), we systematically keep the one with the highest number of parameters. We assign probabilities to the different categories (F, Z, R, Z+R and ZxR), computed as Schwartz BIC weights[241] (Supplementary Table 2.4). All best fits with their parameter values are listed in Supplementary Table 2.1.

2.5.5.6. Comparison of peak times with the *Atger et al.* dataset

We compared our rhythmically classified genes with those obtained from the data in[227]. These data consist of bulk liver RNA-seq data sampled every 2 hours for 24 hours, with 4 replicates per time condition. Thus, there are two main differences between our current and that dataset. First, we can only compare the genes for which rhythmicity is not changing across layers, *viz.* the R and Z+R categories. Secondly, our dataset has a lower temporal resolution, with fewer replicates per time point, meaning that the statistical power

for the detection of rhythmic genes is higher in reference [227]. We therefore did not compare genes found as flat in our dataset.

To assess gene rhythmicity from reference [227], we used harmonic regression on the log-transformed profiles as previously. Using the same notation as above, we define the two following models:

$$\begin{cases} y_{t,i} &= \mu + \varepsilon & (ix) \\ y_{t,i} &= \mu + a\cos(\omega t) + b\sin(\omega t) + \varepsilon & (x) \end{cases}$$

We then fit Eq. $(viii)$ and Eq. (ix) to every transcript, and, for each of them, keep the model with the lowest BIC. Transcripts for which Eq. (ix) is favored are assigned to the flat category (F), while transcripts for which Eq. (x) is favored are assigned to the rhythmic category (R). We then compare the phases and amplitudes of the transcripts classified as rhythmic in both datasets, and compute the circular correlation coefficient[242].

2.5.6. KEGG and DAVID pathway Enrichment analysis

Functional annotation clustering from DAVID[229] for the three categories Z, R, Z+R was ran with standard parameters, using the above set of 5002 selected genes as background set.

We also studied the enrichment of KEGG pathways using hypergeometric tests. In practice, if a background set contains N genes, among which K belong to the category Y, and if a given pathway contains n genes, then the p-value corresponding to an enrichment of k genes is the probability of observing at least k genes from the pathway which belong to the category Y, that is (for a right-sided test):

$$p(x \geq k) = \sum_{x=k}^{n} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad (xi)$$

FDR is then adjusted running the Benjamini-Hochberg procedure on the resulting set of p-values, for each category (zonated, rhythmic, etc.) separately. For clarity, a value z is also computed alongside as:

$$z = \frac{k - \frac{K}{N}n}{\frac{K}{N}n} \quad (xii)$$

This value must be interpreted as the relative difference between the observed number of genes k in the pathway belonging to Y, and the expected value if the distribution was similar as the one from the background.

332 KEGG pathways were tested against KEGG genome T01002. To exclude too specific and too general pathways, pathways having less than 3 or more than 500 genes in common with our dataset were discarded from the analysis.

2.5.7. smFISH

2.5.7.1. Analysis of Z+R and ZxR genes (Stellaris smFIH probes)

Preparation of probe libraries, hybridisation procedure and imaging conditions were previously described[224]. smFISH probe libraries were coupled to TMR, Alexa594 or Cy5. Cell membranes were stained with alexa fluor 488 conjugated phalloidin (Rhenium A12379) that was added to GLOX buffer[243]. Portal node was identified morphologically on DAPI images based on bile ductile, central vein was identified using smFISH for Glul in TMR, included in all hybridisations. For zonation profiles, images were taken as scans spanning the portal node to the central vein. Images were analysed using ImageM[243]. Quantification of zonation profiles in different circadian time point were generated by counting dots and dividing the number of dots in radial layers spanning the porto-central axis by the layer volume. Central vein niche NPCs were identified by co-staining of Aqp1, Igfbp7 and Ptprb. The central vein area was imaged and the images were analysed using ImageM. We counted dots of Wnt2 and Dkk3 expression in NPCs lining the central vein and removed background dots larger than 25 pixels. We then divided the dot count by the segmented cell volume. In total 489 NPCs from 120 central veins of 2 mice (ZT0,6,12,18) were imaged and a Kruskal-Wallis test based on the mean mRNA dot concentration in each cell was performed to compare the timepoints.

2.5.7.2. Temporal analysis of circadian genes (RNA scope smFISH probes)

smFISH of R genes were done on fresh-frozen liver cryosections (8µm) embedded in OCT Compound (Tissue-Tek; Sakura-Finetek USA), sampled every three hours (ZT0 to ZT21). RNAscope® probes for *Bma1l* mRNA (Mm-Arntl, catalog #: 438748-C3) and *Per1* mRNA (Mm-Per1, catalog #: 438751) were used, according to the manufacturer's instructions for the RNAscope Fluorescent Multiplex V1 Assay (Advanced Cell Diagnostics). To detect the central vein, an immunofluorescence of Glutamine Synthetase (ab49873, Abcam, diluted 1:2000 in PBS/BSA 0.5%/Triton-X0.01%) was done together with smFISH. Nuclei were counterstained with DAPI and sections were mounted with ProLong™ Gold Antifade Mountant. Liver sections were imaged with a Leica DM5500 widefield microscope and an oil-immersion x63 objective. Z-stacks were acquired (0.2µm between each Z position) and mRNA transcripts were quantified using ImageJ, as described previously in reference [232].

Pericentral (PC) and Periportal (PP) veins were manually detected based on Glutamine Synthetase IF or on bile ducts (DAPI staining). The Euclidean distance between two veins and the distance from the vein of each mRNA transcript were calculated. mRNA transcripts were assigned to a PP or PC zone if the distance from the corresponding vein was smaller than one-third of the distance between the PP and PC veins (ranging from 50 to 130μm).

2.5.7.3. Wnt2 and Dkk3 expression in LEC

Preparation of probe libraries, hybridisation protocol and imaging conditions were previously described[224]. The *Aqp1, Igfbp7 and Ptprb* probe libraries were coupled to TMR, the *Wnt2* library was coupled to *Cy3* and the *Dkk3* library was coupled to *Cy5*. Cell membranes were stained with alexa fluor 488 coupled to phalloidin (Rhenium A12379) that was added to GLOX buffer[243]. The central vein was identified based on morphological features inspected in the DAPI and Phalloidin channels and presence of Wnt2-mRNA (detected by smFISH). Endothelial cells were identified by co-staining of *Aqp1*, *Igfbp7* and *Ptprb*. The central vein area was imaged and the images were analysed using ImageM[243]. We counted dots of *Wnt2* and *Dkk3* expression (corresponding to single mRNA molecules) in endothelial cells lining the central vein and removed background dots larger than 25 pixel3. We then divided the dot count by the segmented cell volume. In total 516 endothelial cells from 120 central veins of at least 2 mice per time point.

2.5.8. Data availability

2.5.8.1. scRNA-seq data

All scRNA-seq data is deposited in GEO with accession code GSE145197 (reviewer token ezobmqmqftcrpeh).

2.5.8.2. Reconstructed gene profiles

Reconstructed spatio-temporal gene profiles are available as Matlab files at https://c4science.ch/diffusion/10261/

2.5.9. Web-application

The whole dataset of gene profiles along with the analysis is available online as a web-application at the URL: https://czviz.epfl.ch/. The application was built in Python using the library *Dash by Plotly* (version 1.0).

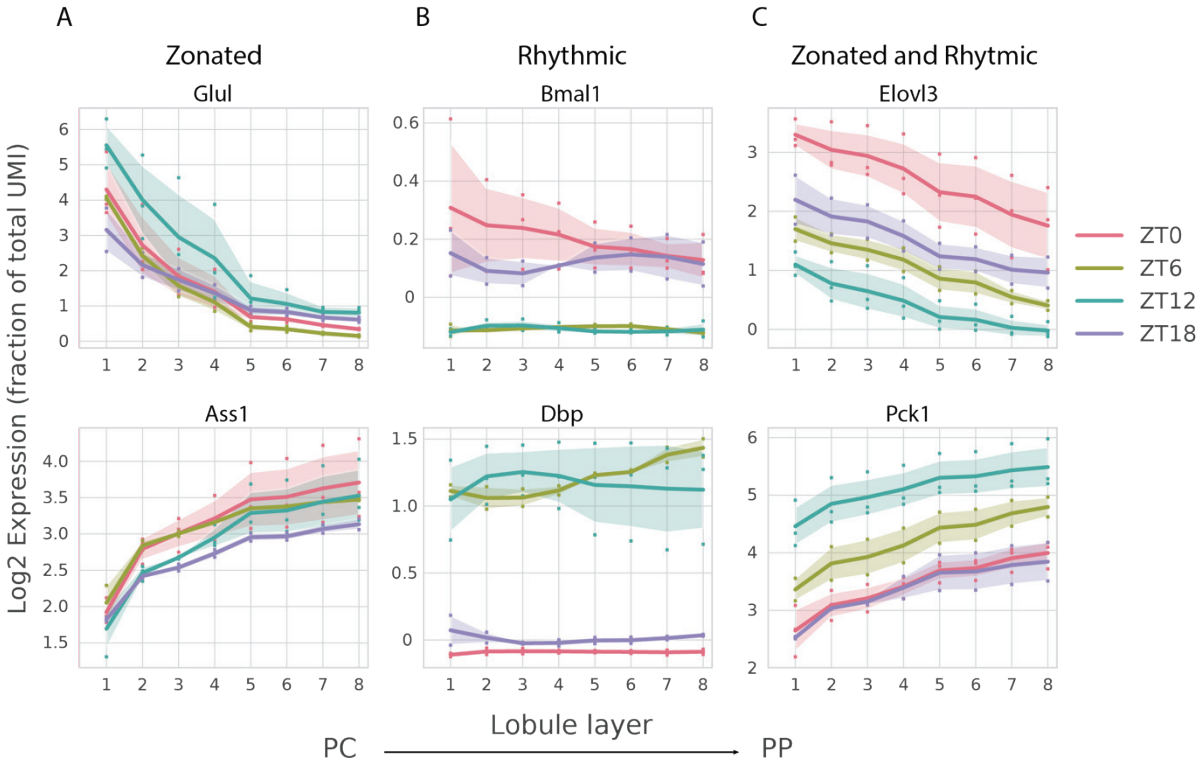
2.5.10. Code availability

The code for fitting the mixed-effects models and generating the main figures is available at https://c4science.ch/diffusion/10261/

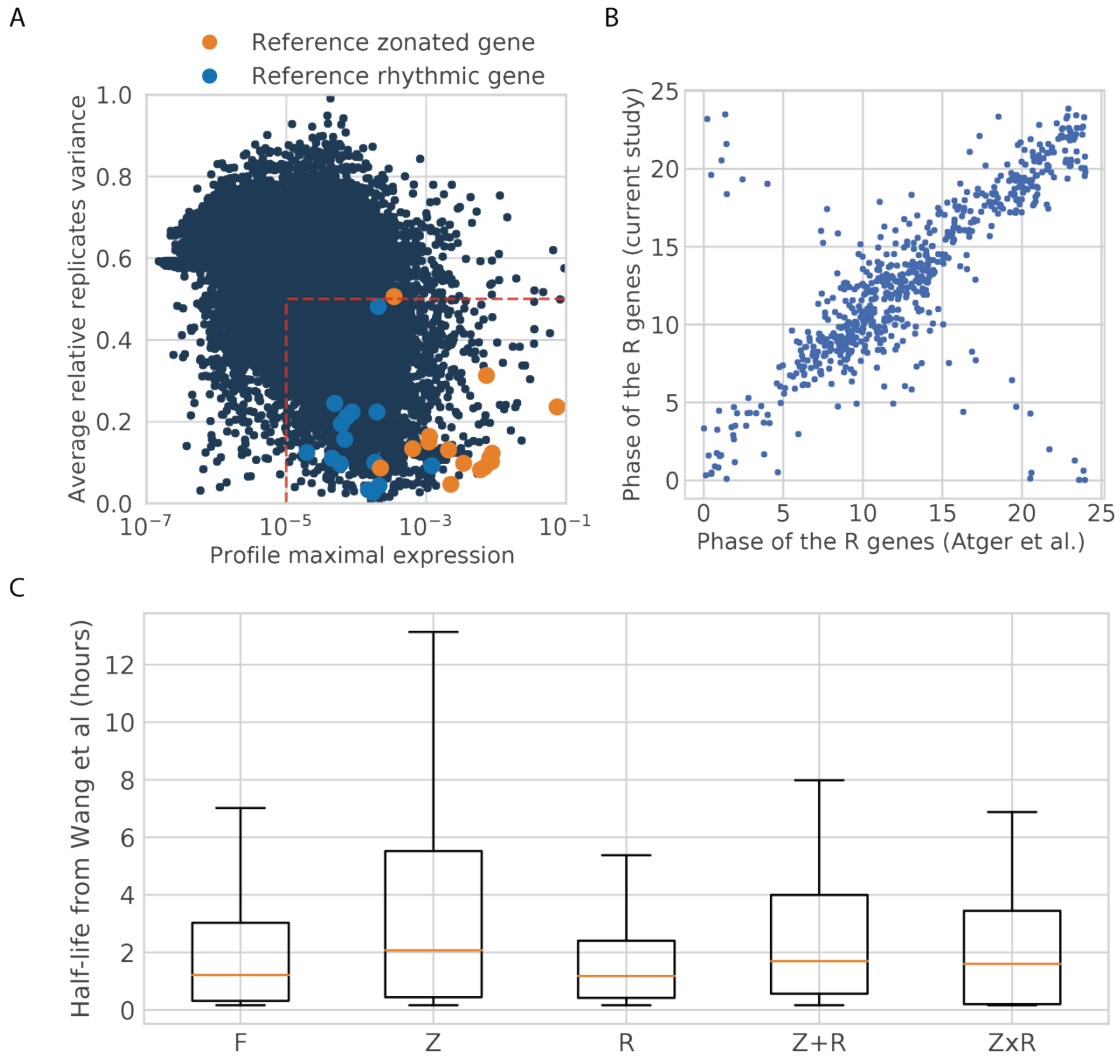
2.6. References

For consistency, references from the article have been placed at the end of the thesis

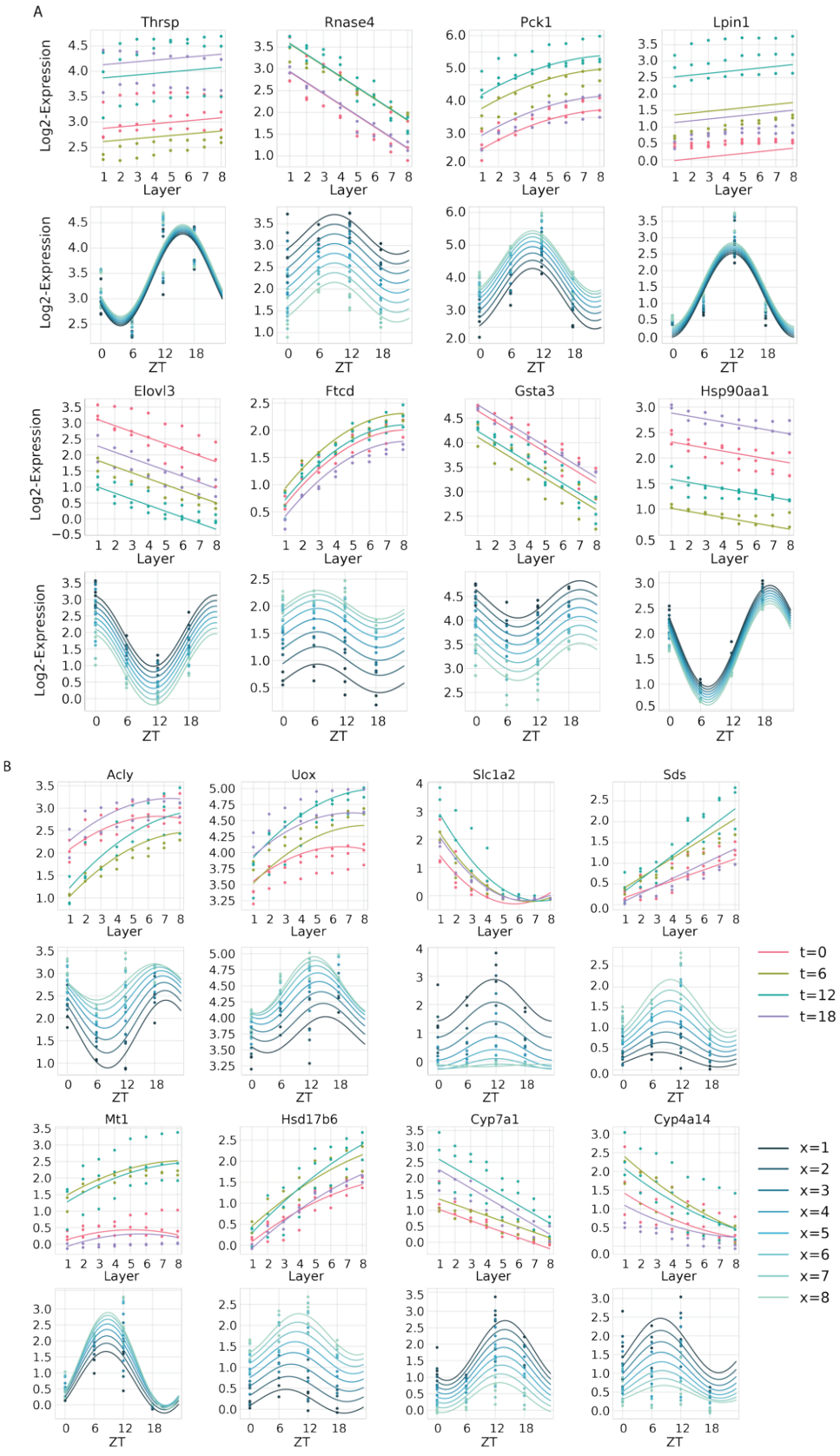
2.7. Supplementary Figures



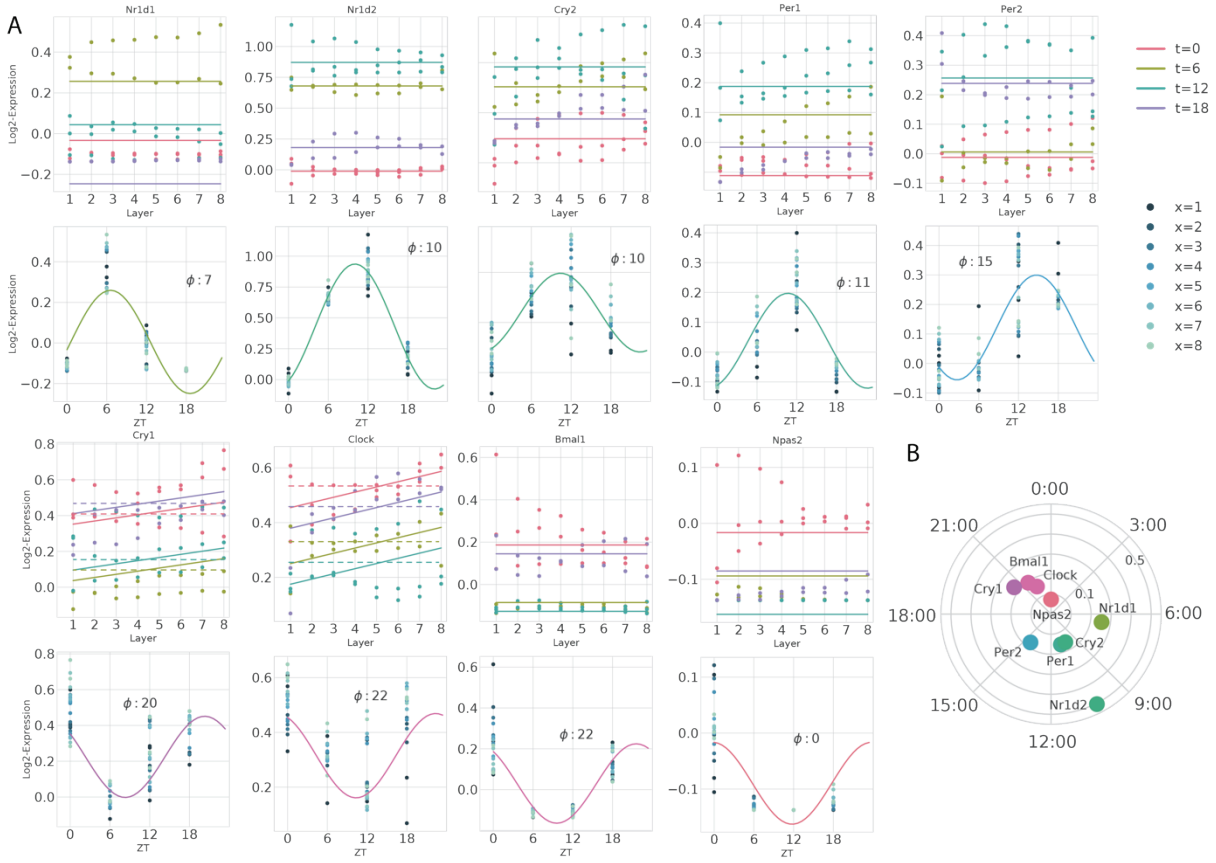
Supplementary Figure 2.1: Log-transformed expression profiles. (**A-C**) Expression levels of the reconstructed profiles for the example genes from Figure 2.1, panels F-H after log-transformation (Methods). Shaded areas represent SD across mice.



Supplementary Figure 2.2: Pre-filtering of the genes and comparison with external datasets. (**A**) Biological variability of gene profiles across independent replicate liver samples, quantified in terms of the average relative replicate variance. **0** shows perfectly reproducible profiles while **1** the most variable genes (Methods). Gene inside the bottom-right box (x-cutoff at 10^{-5}; y-cutoff at **0.5**) are selected and contain all but one of the reference genes. Colored dots show reference zonated genes (blue) and reference rhythmic genes (orange). (**B**) Comparison of the peak time for rhythmic genes in R and Z+R, with the dataset from *Atger et al, 2015*[227]. Circular correlation coefficient is 0.746 (Methods). (**C**) Boxplot of the mRNA half-lives (data from *Wang, J. et al, 2017*[223]) shows that R genes as a group (median, orange line) are the shortest-lived. Box limits are lower and upper quartile, whiskers extend up to the first datum greater/lower than the upper/lower quartile plus 1.5 times the interquartile range. Remaining points are outliers.



Supplementary Figure 2.3 (modified version, verticalized): Spatial and temporal profiles of the outliers from Figure 2.3C-D. (**A**) Spatial (top) and temporal (bottom) representation of eight Z+R genes showing large slope and/or amplitude in Figure 2.3C. (**B**) Same representations with eight ZxR outliers selected from Figure 2.3D having volatile phase and/or amplitude.



Supplementary Figure 2.4: Core-clock circadian genes escape zonation. (**A**) Spatial and temporal profiles and fits for circadian core-clock genes. Peak times are indicated on the temporal representation. For the genes *Cry1* and *Clock*, additional dashed lines represent fits for the R model, as the Schwartz BIC weights from the R and Z+R models were close (Supplementary Table 2.4). (**B**) Amplitudes and peak times of the core-clock circadian genes in a polar coordinate representation (clock-wise ZT times are indicated, distance from the center corresponds to the amplitude) show the expected organization of core clock transcript in the liver.

Supplementary Figure 2.5: Space-time logic of signaling pathways. Enrichment/depletion of genes targeted by the Wnt, Ras and Hypoxia pathways with respect to the main background sets: all genes (column B1 and B5), R and Z+R (B2), Z+R (B3), and Z and Z+R (B4). Tested sets are rhythmic genes (R and Z+R, B1), diurnally and nocturnally rhythmic genes ($R^{night/day}$ and $Z+R^{night/day}$, B2), diurnally/nocturnally central/portal genes ($Z^{c/p}+R^{night/day}$, B3), central and portal genes ($Z^{c/p}$ and $Z^{c/p}+R$, B4), and zonated genes (Z and Z+R, B5). Example of interpretation: the positive targets of Ras are not enriched nor depleted in rhythmic genes, but they are highly enriched in zonated genes. Among the targets that are zonated, we observe an over-representation of portal genes with respect to what is observed in the background of all the (filtered) zonated genes; accordingly, we observe a under-representation of positive central targets. Moreover, the positive targets of Ras which are rhythmic and portal appear to peak preferentially at night (compared to the distribution of peak times of all the zonated-rhythmic genes).

2.8. Supplementary Tables

Supplementary Tables are not adapted to the format of this thesis but are available online along with the preprint version of this study.

3. Web application

3.1. Introduction

In the project presented Section 2, we're dealing with a dataset of about fifteen thousand genes, each of which being represented by a three-dimensional profile of eighty datapoints. The corresponding analysis involves ten times as much data, as the resolution for the fits is much higher than the original profiles. As it would be cumbersome to share the whole dataset and analysis as Excel files, we decided to store everything in a database like-format, accessible directly online. To simplify even more the user interface, we coded a Python web-application using the library Dash by Plotly, enabling to explore the different gene profiles and fits from the browser.

The application, called CZViz, is fully working and available at https://czviz.epfl.ch/.

3.2. What is a web-app?

Data can be cumbersome to explore, especially when heavy. Similarly, analysis results can be hard to interpret from a CSV file or an Excel Sheet. Web applications (Web-app), as a client-server computer program, enable the consultation and download of datasets and analysis using an intuitive Graphical User Interface (GUI) directly in a browser. Common web-apps include webmail or online banking and, as such, are already used by a majority of the adult population[244].

Early web-apps were excessively slow as the application computations were shared between the server the client, posing compatibility problem and adding to the support cost, as an upgrade from the server also needed an upgrade from the client. Besides, each web page was shown as a static document, and interactivity was only simulated as a sequence of pages, such that any update made to one page would require the server the reload the whole sequence.

With the development of HTML5 and Javascript, modern web-apps run only on the server-side (except if not needed), and are supported by a wide variety of internet browsers. Web applications can be considered as a specific variant of client-server software where the client software is downloaded to the client machine when visiting the relevant web page, using standard procedures such as HTTP. Client web software updates may happen each time the web page is visited. During the session, the web browser interprets and displays the pages, and acts as the universal client for any web application[244].

Nowadays, web-apps can be coded directly from "classic" programming languages such as Python and R, using libraries developed for the purpose[245], [246]. These libraries usually work as a wrapper for javascript/HTML5 code, but in practice, they deliver an experience almost as fast as a natively implemented application.

3.3. CZViz

Figure 2.7: GUI of CZViz (main page, gene selection). The gene of interest can be selected using the left dropdown menu, which is then represented under various perspectives on the right.

CZViz is a Dash web-application, itself embedded in a website coded using HTML 5 and Bootstrap 4. The structure of the application is as follow:

- Tabulation 1: Gene selection (Figure 2.7)
    - Left: Introduction to the application, followed by a control card to select the gene of interest (a search engine is integrated)
    - Right: Four panels, each corresponding to a different way of presenting the data: either in space, time, both, or the raw data in a table.
- Tabulation 2: Statistical analysis (Figure 2.8)
    - Left: type of analysis. Choices are "Spatial analysis", "Temporal analysis", "Spatiotemporal analysis" or "Validation".

- Right: results from the current analysis, representing visually and in tables the different types of regressions.

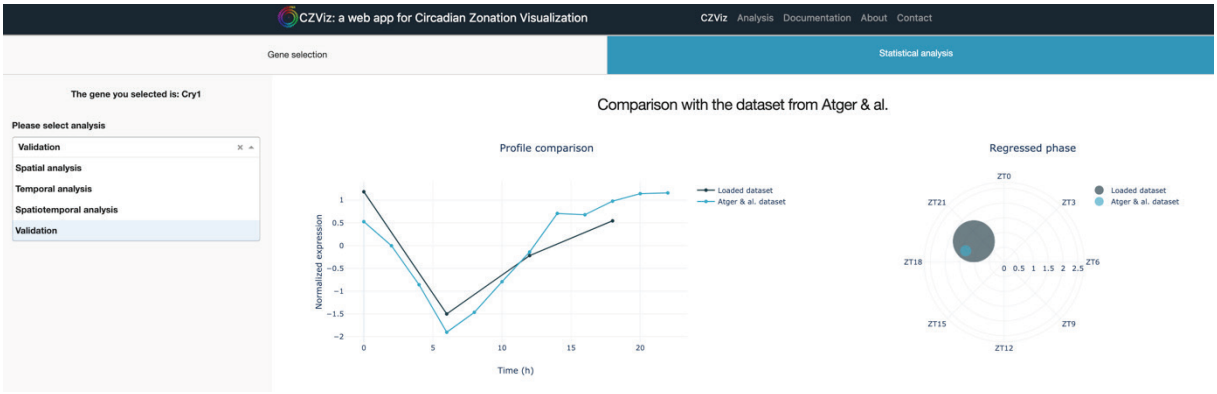


Figure 2.8: GUI of CZViz (secondary page, statistical analysis). The type of analysis can be selected using the left dropdown menu. The corresponding analysis is returned on the right, with the appropriate representation.

For more details, please refer to the app documentation. Due to memory limitation, all analyses are done on the fly on the lab server, using Python scripts running with Gunicorn. Most analyses take less than a second to run.

3.4. Code availability

The corresponding Python code is open-source, available on GitHub: https://github.com/ColasDroin/CZViz

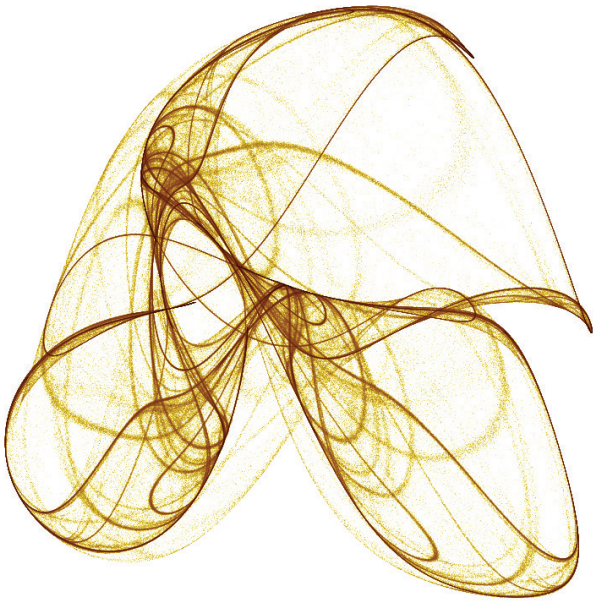
Chapter 3: RNA velocity-based inference of cell cycle properties using single-cells

This work is currently being written, for submission as a method paper. Authors list is as follow (definitive order not decided): L. Talamanca, C. Droin, A. Lederer, Giolele La Manno, F. Naef.

Contributions

In all the tasks listed below, my work is always under the supervision of F. Naef.

F. Naef, G. La Manno and L. Talamanca initially designed the study, with some changes following my involvement a few weeks later. G. La Manno and A. Lederer collected and partly wrangled the datasets. Along with the help of F. Naef, G. La Manno and I, L. Talamanca designed the inference method. I designed the simulation method and implemented the inference method. With the help of G. La Manno, I designed and implemented the corresponding parameter optimization method. I ran the inference method on simulated and real data. All authors are currently involved in the writing of the manuscript.



Artwork Figure 4: Artistic representation of the De Jong strange attractor. On each of its three dimensions, this attractor generates oscillations, whose amplitude and period are never exactly the same from cycle to cycle. Simulated with Processing, based on the work of Robert D'Arcy.

1. Project Introduction

1.1. Motivation and aims

Breakthrough methods in single-cell transcriptomics have recently opened new opportunities for the study of dynamic biological processes such as cell development and differentiation[247], [248]. Among them, unsupervised trajectory inference algorithms aim to reconstruct in an unbiased manner the underlying dynamical processes that occur in a cell population[249].

However, to get a genuinely predictive trajectory model, temporal information is required, at least to constrain the space of possible dynamics[250]. This is precisely what RNA velocity has brought, by involving the fact that nascent mRNAs and spliced mRNAs, which are causally linked (the first "predicts" the second at a latter time point), are actually distinguishable in many scRNA-seq protocols, due to the presence of introns in the formers[251]. Therefore, by building a per-gene reaction model (production-degradation rate model) describing how unspliced mRNA is progressively transformed into spliced mRNA before being degraded, changes in mature mRNA abundance can be predicted. By combining these predictions across all genes, one can estimate towards which state a single cell is evolving.

In its most accurate form, RNA velocity needs to incorporate the mRNA splicing and degradation rates for each transcript. Yet, in a standard single RNA-seq snapshot dataset, only the ratio of the two is easily accessible, assumiong moreover that the genes under study reach steady-state. Therefore, a significant hypothesis has been made: the splicing rate is taken as common to all genes[251].

In this project, we wanted to revisit RNA velocity inference such that the obtained vector fields would bear dynamical significance. To this end, we decided to exploit that in the high-dimensional gene space, cell velocities are forced to remain tangent to these manifolds. Adding such a hypothesis helps to handle the noise (both intrinsic and extrinsic), as all cells must be projected onto the manifold but, due to the simpler structure of the model.

We decided to start with the cell-cycle, as it necessarily follows a limit-cycle in expression space, and is already well described in the literature. Corresponding biological questions that we aim to answer are: how does the speed of the cell-cycle depend on the phase of the cell? Does that change with the cell-type or growth conditions? Does that change with external factors, such as temperature?

The difficulty to answer these questions is that the model is hard to optimize, as it involves many parameters and quite noisy data.

1.2. Background

1.2.1. Single-cell temporal-omics[23]

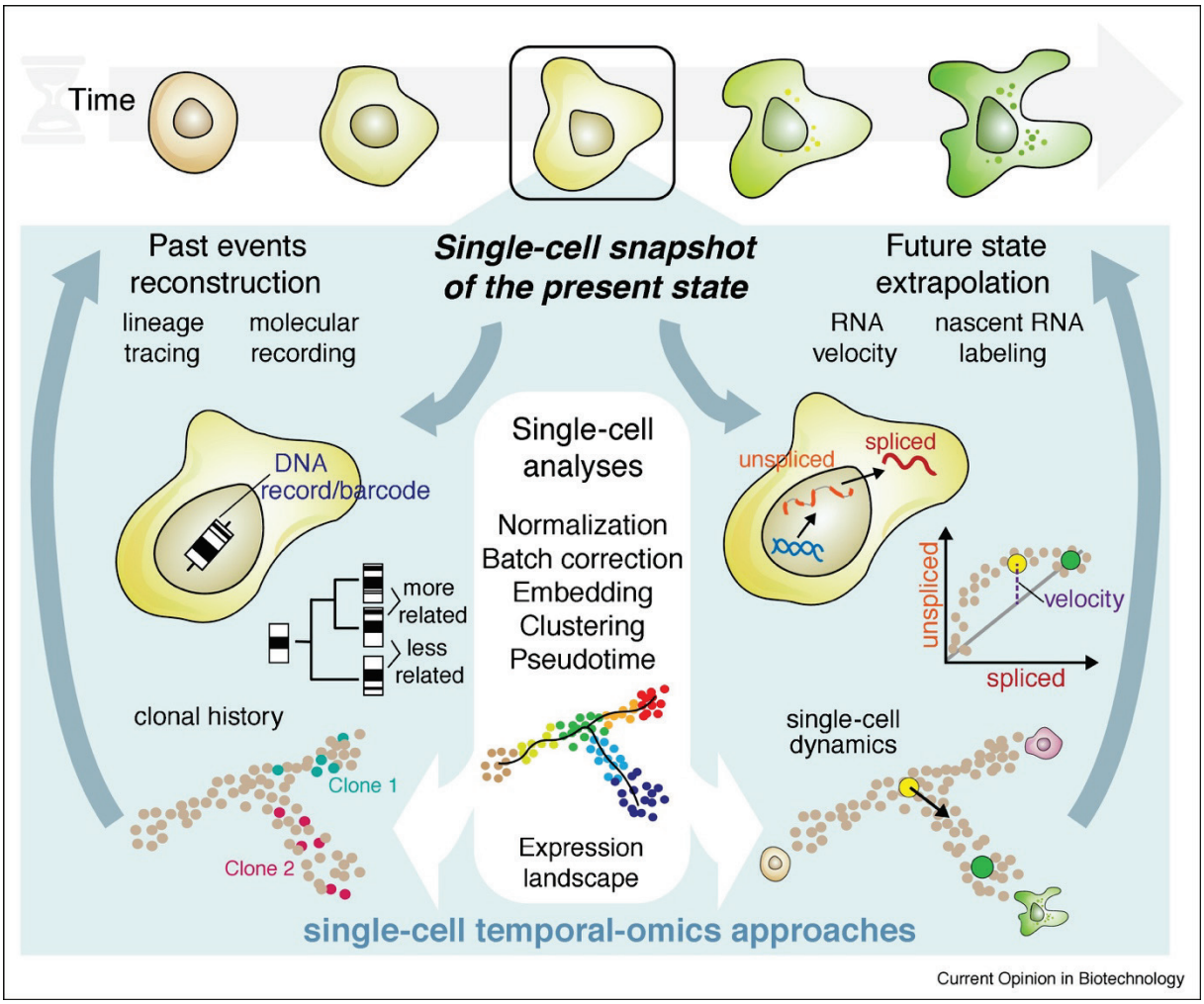
1.2.1.1. Introduction

Single-cell transcriptomics is a relatively recent field of science[253] which takes an interest in analyzing gene expression in various biological systems at the resolution of individual cells. To this end, it highly relies on Single-cell RNA sequencing (scRNA-seq), a technology that has been used in very different contexts: building tissue atlases[254], [255], enhance genome-wide association studies perspectives[256], unravel the complexity of developmental processes and gene regulatory networks, or build cell fate commitment landscapes[257].

When applied to a population cells (e.g. dissociated from a tissue samples, or cells grown in culture), scRNA-seq can provide a static snapshot of the cell states, containing information about the biological processes they undergo. For instance, if cells are differentiating, gene expression space will be spannning mature and transient cell types. However, contrary to live cell approaches, snapshot scRNA-seq data does not infer temporal cellular trajectorie, but rather reveals variation in expression among cell types within the cells at the population level.

Now, modern scRNA-seq techniques are very high-throughput[258], [259], but the interpretation of the information they provide into biological knowledge is not an easy task, usually requiring dedicated statistical methods and modelling approaches[253], [260]. Among them, many have been developed to assess dynamical or regulatory aspects of the data: RNA velocity[251] is probably the most famous (and the most relevant to this work), but this also includes nascent RNA quantification[261], lineage tracing[262] or molecular recording[263]. A short explanation of single-cell temporal omics approaches is provided in Background Figure 3.1.

[23] This introduction to single-cell temporal-omics is partly adapted from a review[252] by two co-authors of the study presented Section 2: Alex Lederer and Gioele La Manno.



Background Figure 3.1: Illustrative representation of single-cell temporal-omics approaches. Given the snapshot obtained from single-cell RNA sequencing, data analyses enable characterization and classification of the gene expression landscape in a heterogeneous population of cells. Recent methods further enable the extrapolation of future gene expression states (**right**) and reconstruction of past cellular events (**left**). Together, these approaches permit greater inference of the temporal changes within a single cell while still relying on measurement from a single time point. Figure and caption taken from *Lederer et La Manno, 2020*[252].

1.2.1.2. Pseudotime trajectories

Developmental biologists are often interested in following the time trajectories of single-cells in gene expression space, as this explains how a cell can transition from one state to another. Unfortunately, repeated extraction and sequencing of cellular contents from living cells is not an efficient option yet[264]. As of now, most single-cell sequencing techniques will kill the cells being sequenced, and, therefore, can only provide a static snapshot of the cell-state at the time of sampling[250]. However, assuming that the sequenced cells follow the same trajectories[24], it is possible to order them in a way that captures internal time. Methods

[24] And that a sample contains sufficient heterogeneity in causally related cellular states.

such as Monocle[265], Wanderlust[266] or Cyclum[267] have been developed for this exact purpose, introducing the concept of "pseudotime" as a proxy for internal time. As of 2020, more than 70 pseudotime inference methods have been published in the literature[267].

Trajectory inference methods have several drawbacks, and the main one is probably that they poorly handle variations in cell density in the expression space. For instance, if cells tend to accumulate towards a stable sate, the corresponding transient states may be distorted. Therefore, pseudotime analysis is not and should not be expected to reveal the real trajectories of the cells. Still, this doesn't prevent it from bringing valuable biological insights[268].

In 2017, a method called "RNA velocity" was developed by *La Manno et al.*[251] to precisely address this drawback, by involving dynamical information in the model and estimation process.

1.2.1.3. RNA velocity

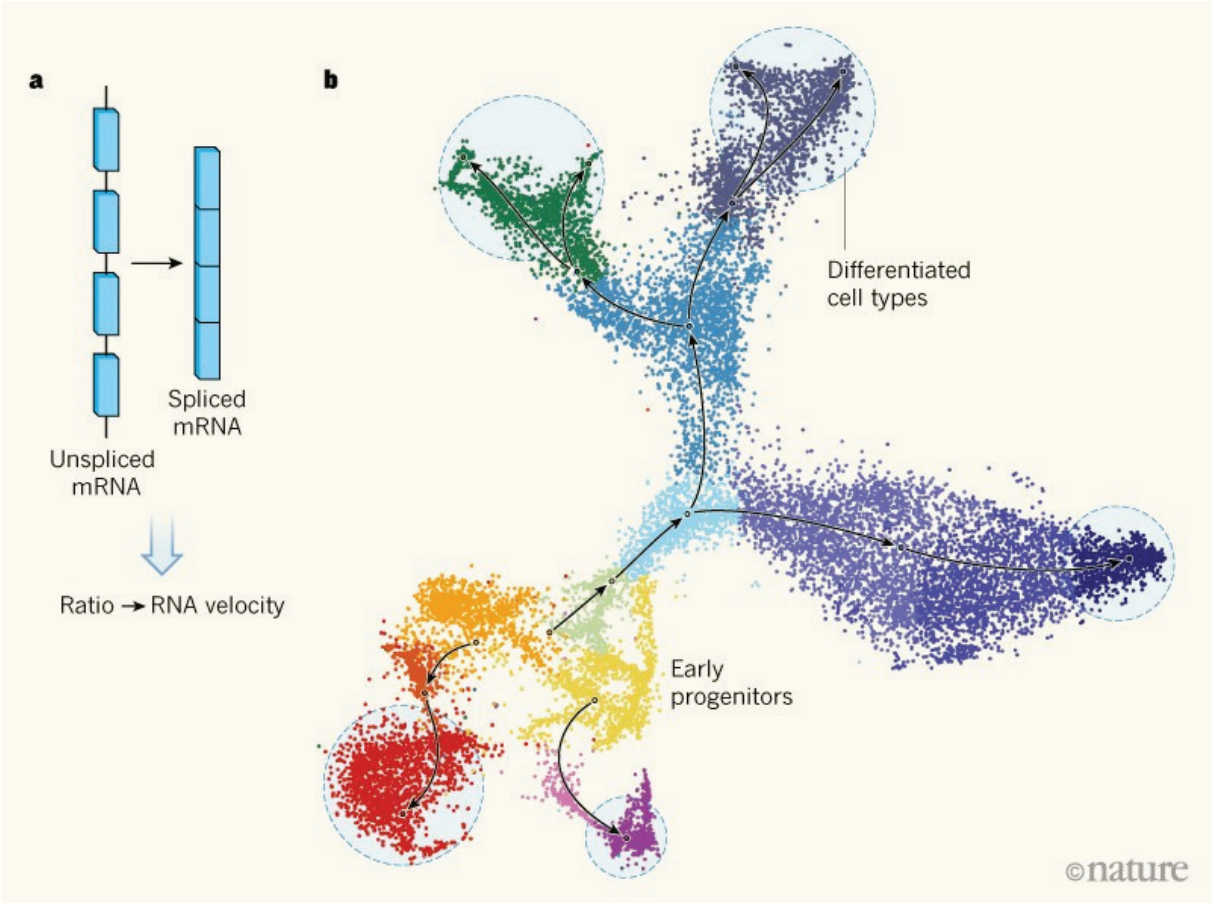
RNA velocity takes advantage of the fact that the production of mRNA is progressive: the nascent mRNA is transcribed from the DNA, before getting spliced and ready to get translated into proteins. Therefore, the amount of unspliced mRNA will necessarily determine the amount of spliced mRNA, such that a rate equation can be found which describe how these mRNA quantities change in time. In practice, RNA velocity uses intronic reads as a proxy for unspliced mRNA levels in scRNA-seq protocols, while exonic reads for spliced mRNA levels. The corresponding system of equations describes the mRNA dynamics[251]:

$$\begin{cases} \frac{du}{dt} = \alpha - u(t) \\ \frac{ds}{dt} = u(t) - \gamma s(t) \end{cases} \qquad (i)$$

Here, u and s stand for the number of unspliced and spliced mRNA, respectively. α is the transcription rate, assumed constant, and γ is the degradation rate. A central hypothesis of this model is that the splicing rate is constant, equal to 1. This is not so much of a problem in itself for one gene as it just means that the resulting mRNA counts are given in units of splicing. However, applying this model to several genes at the same time, as it is done, is equivalent to assuming that all genes have the same splicing rate, which is most likely wrong[269], [270].

Eq (i) is a first order a differential equation describing how mRNA expression is evolving in time. Knowing u and s at a given time point allows for the prediction of mRNA expression at future times for all genes of a given cell. One can thus have an idea of the direction it is going to take in the expression space, that is, what is going to be its real temporal trajectory

(Background Figure 3.2). RNA velocity has already been applied with success in developmental studies[271], [272] and diseases[273], [274]. Now, mainly because of the strong assumptions regarding the constant transcription and splicing rates, this method also has drawbacks and remains imprecise.



Background Figure 3.2: Measuring dynamic changes in gene expression across complex tissues. (**a**) As messenger RNA matures, sections of the immature transcript are removed — a process called splicing. When the expression of a gene increases, a transient increase in the proportion of immature, unspliced transcripts compared with that of mature, spliced transcripts is observed in the cell. By contrast, a higher proportion of spliced transcripts is seen for a short time when expression of the gene decreases (not shown). La Manno *et al.*[251] measured the ratio of unspliced to spliced transcripts for each gene in a single cell to calculate a quantity called the RNA velocity, which reveals how gene expression is changing. (**b**) By measuring RNA velocity in thousands of cells in a tissue (here, in neurons in the developing mouse brain), the authors could generate maps that show not only how closely related cells are to one another (with closeness indicated by similar colours), but also which cells they will become similar to in the future (indicated by arrows), according to the gene-expression changes they are undergoing. RNA velocity successfully tracks early progenitors (orange and yellow) that eventually give rise to a range of differentiated cell types (blue dashed circles). Figure and caption taken from the *News and Views* article by *Allon Klein*[275].

1.2.2. Optimization: L-BFGS-B

In this project, parameter optimization is performed using L-BFGS-B, also known as limited memory BFGS. In practice, L-BFGS-B is an approximation of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, which enables to solve non-linear unconstrained optimization problems. BFGS is itself a quasi-newtonian method, that is, a method that looks for the zeros of a function whose Jacobian or Hessian analytical expression is not known[276].

The main idea of BFGS is to avoid to explicitly building the Hessian matrix, and instead make an approximation of the second derivative inverse of the function to optimize, by analyzing successive gradients. This approximation relies on the assumption that the function to optimize can be locally approximated by a quadratic Taylor expansion around the optimum:

$$f(\boldsymbol{x}+\boldsymbol{\varepsilon}) \approx q(\boldsymbol{\varepsilon}) = f(\boldsymbol{x}) + \boldsymbol{\varepsilon}^T g(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\varepsilon}^T H(\boldsymbol{x})\boldsymbol{\varepsilon} \qquad (ii)$$

Here, f is the function to optimize, $\boldsymbol{g(x)}$ is the gradient vector and $\boldsymbol{H(x)}$ is the Hessian matrix**.** By taking the derivative of this expression, it can be shown that the necessary condition for a local minimum of $\boldsymbol{q(\varepsilon)}$ results in the following linear system:

$$g(\boldsymbol{x}) + H(\boldsymbol{x})\boldsymbol{\varepsilon} = 0 \qquad (iii)$$

In turn, this expression leads to the Newton direction $\boldsymbol{\varepsilon}$ for line search:

$$\boldsymbol{\varepsilon} = -H(\boldsymbol{x})^{-1} g(\boldsymbol{x}) \qquad (iv)$$

In practice, the Newton direction is only reliable if, on the one hand, the Hessian matrix exists and is positive definite, and on the other hand, the quadratic approximation is reasonably good.

The approximation of the Hessian matrix $\boldsymbol{B(x)} \approx \boldsymbol{H(x)}$ is done with an update formula called the BFGS updating formula:

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k - \frac{\boldsymbol{B}_k \boldsymbol{s}_k \boldsymbol{s}_k^t \boldsymbol{B}_k}{\boldsymbol{s}_k^t \boldsymbol{B}_k \boldsymbol{s}_k} + \frac{\boldsymbol{y}_k \boldsymbol{y}_k^T}{\boldsymbol{y}_k^T \boldsymbol{s}_k} \qquad (v)$$

In this equation, $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ and $\boldsymbol{y}_k = \boldsymbol{g}_{k+1} - \boldsymbol{g}_k$, while the vectors $\boldsymbol{B}_k$ and $\boldsymbol{g}_k$ are approximations of the functions B and g at iteration k. The initial condition $\boldsymbol{B}_0$ can be taken as any symmetric positive definite matrix (e.g. the identity matrix).

The BFGS updating formula can be shown to converge to $H(\boldsymbol{x}^*)$, where $\boldsymbol{x}^*$ is f (local) optimum, with superlinear convergence[277].

2. Study in preparation for publication

2.1. Abstract

The cell-cycle is one of the main drivers of gene expression cell-to-cell heterogeneity in otherwise homogeneous cell populations. Although various methods have been developed to characterize its progression, they usually rely on a few known markers. scRNA-seq provides a static snapshot of gene expression levels across a cell population. While the gene expression between cell types varies across multiple axes of variation, cellular processes such as the cell-cycle tend to unwind on lower-dimensional manifolds. RNA velocity analysis introduced the inference of vector fields in gene expression space, opening the way towards temporal interpretations of biological processes from scRNA-seq. Here we develop this idea further by formulating the problem in terms of an autonomous dynamical system and use this to infer a consistent model for cell-cycle dynamics on the circle. The corresponding cell phases on the circular manifold are identified using an Expectation-Maximization-based phase inference method. We validate our approach on several scRNA-seq datasets, revealing distinct proliferation modes in different cell-types.

2.2. Introduction

Until recently, transcriptome studies were done using bulk tissues, considering different cells as homogeneous objects and thereby ignoring intrinsic variability of gene expression. By examining gene expression level of individual cells, single-cell transcriptomics has allowed for the simultaneous measurement of mRNAs from thousands of gene [278]. This means that biological processes such as cellular differentiation and lineage choice can now be studied in a high-throughput and unbiased manner. This relatively new field of developmental biology is called trajectory inference[279].

Given a heterogeneous cell population, trajectory inference aims to find out which transcriptional changes led to which cell state, and how. To that end, many different computational and mathematical methods have been developed, almost all considering cell dynamics on a tree of branching trajectories, in a low-dimensional space[249]. These methods can be powerful but are limited by the structure of the single-cell RNA-seq data, which only contains a static snapshot of cellular states.

Ideally, one would like to follow the evolution of individual cells in time, but this is currently hard to do with contemporary techniques[280]. To get a truly predictive trajectory model, temporal information seems required to constrain the space of possible dynamics[250]. RNA velocity analysis has made significant steps in this direction, by exploiting the fact that the

production of mRNA is progressive: for any given gene, the nascent mRNA is transcribed from the DNA, before getting spliced and prepared for translation into proteins. Therefore, the amount of unspliced mRNA will determine the amount of spliced mRNA, such that a per-gene rate equation can be found which describes how these quantities change in time. By integrating these rate equations across all genes, one can estimate towards which state a single cell is evolving. RNA velocity does not require the development of new assays, as nascent mRNAs and spliced mRNAs are both captured by most scRNA-seq protocols[251].

The differential equation presented in the original RNA velocity paper require the estimation of the splicing and degradation rates to make efficient predictions. Yet, in the data, only the ratio of the two is easily accessible, and even so it needs to be estimated in a heuristics ways by quantile regression. As a first approximation, the foundational paper by *La Manno et al.* makes the a strong hypothesis that the splicing rate is the same for all genes[251].

We here present a method that deals with this issue by constraining the corresponding dynamical model. Indeed, although the dimensionality of cellular state variables is usually large, most cellular processes actually unwind on low-dimensional manifolds. By assuming that the cells velocities are always tangent to such manifolds, we greatly simplify the structure of the underlying dynamics. Adding such a hypothesis not only helps to handle both intrinsic and technical noise but also enables here to infer the effective splicing and degradation rate, although the global scale remains free.

In this introductory paper, we decided to study a circular manifold, the cell-cycle, although our method could be adapted to any well-parametrized manifold. The cell-cycle is particularly interesting as it is well described in the literature[103], and is known to be an important driver of gene expression cell-to-cell heterogeneity in otherwise homogeneous cell populations[281]. The corresponding manifold is topologically a circle, as the cell-cycle can be seen as a closed, periodic, trajectory in expression space[282]. To this day, several methods have been developed to characterize cell-cycle progression from transcriptomics data, but they usually rely on a few known markers and often suffer batch effects[267].

To this end, we adapted a soon-to-be published circadian time inference method by *Talamanca et al.* to handle cell-cycle phases. We then describe how the molecular speed of the cell-cycle depends on the cell phase. We validate our method on several mice and human scRNA-seq datasets, revealing distinct proliferation modes in different cell-types.

2.3. Method

2.3.1. A maximum-likelihood approach to infer cell cycle properties using single-cells

Hereafter, we write the dynamical system governing spliced and unspliced mRNA counts. Our model assumes that the system evolves on a low dimensional manifold embedded in the high dimensional gene space from which we draw the measurements. Accordingly, we develop a probabilistic model for the spliced and unspliced mRNA counts based on the dynamical system. We then integrate out some of the parameters so that we can find an optimizable likelihood function.

2.3.1.1. Geometry of the problem and fundamental assumptions

Our data consist of two measurements per gene per cell: the unspliced and spliced counts. We only use the spliced counts to represent the cell position in expression space. Therefore, each cell is a point in $\mathbb{R}^G$ where G is the number of genes. The unspliced information is used later to build a vector space that faithfully represents the RNA dynamics ocurring in single cells.

The central model assumption is that, in a noiseless version of the system, all cells lie on a low dimensional manifold embedded in $\mathbb{R}^G$, and evolve along trajectories which are also on the manifold (Figure 3.1). Let $\mathcal{M}$ be the manifold and x the corresponding cell coordinates. The predicted gene expression vector for gene g from the position of the cell c in the low dimensional manifold is $y_g(x^c) = (u_g(x^c), s_g(x^c)) = (u_g^c, s_g^c)$, while the measured one is $Y_g^c = (U_g^c, S_g^c)$. Calling β_g the splicing efficiency and γ_g the degradation rate for gene g, expressed in h^{-1}, the time evolution of the spliced counts is governed by:

$$\frac{ds_g}{dt} = F(y_g) = \beta_g u_g - \gamma_g s_g \qquad (1)$$

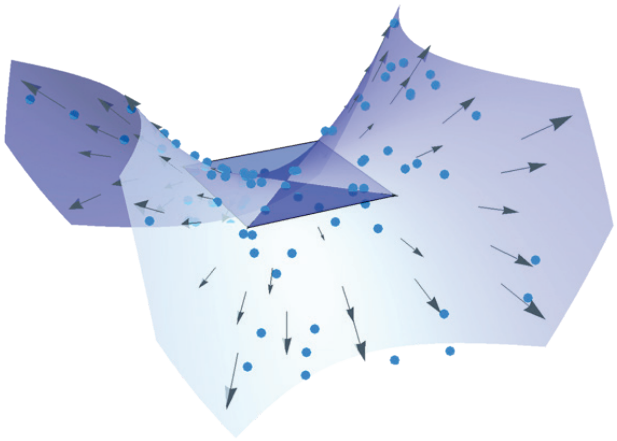


Figure 3.1: Schematic representation of a simple manifold (light blue surface) in a three-dimensional space. Single cells are represented as individual blue points, evolving on the manifold according to the vectorfield represented with grey arrows. The tangent space to the manifold, chosen here for a cell at the center of the saddle-point, is represented as a dark blue square, partially going through the manifold on both of its ascending directions.

Note that both β_g and γ_g are considered time-independent, while the time evolution on $\mathcal{M}$ is deterministic. Using the chain rule for differentiation reveals the following relationship for the spliced counts:

$$\frac{ds_g(x)}{dt} = \nabla_x s_g(x) \cdot \frac{dx}{dt} \qquad (2)$$

The right-hand side of this equation is comprised of two conceptually distinct contributions. One is the mapping between the low dimensional manifold coordinates and the cartesian coordinates of the high dimensional gene space; in particular, it is the first-order approximation of the mapping around the point of interest. The second contribution is the time evolution of the system on the manifold. In particular, in our model, this is governed by an autonomous equation:

$$\frac{dx}{dt} = v(x) \qquad (3)$$

We are interested in reconstructing $v(x)$, which can be interpreted as a velocity vector field on $\mathcal{M}$. Summing up, we get a consistency equation for our model:

$$\nabla_x s_g(x) \cdot \frac{dx}{dt} = \frac{ds_g(x)}{dt} = \beta_g u_g(x) - \gamma_g s_g(x) \qquad (4)$$

We would like $v(x)$ to be directly interpretable, meaning that the trajectories on $\mathcal{M}$ dictated by $v(x)$ should be, when mapped back to gene space, the trajectories that cells follow during a differentiation process.

2.3.1.2. Manifold properties

As a geometric object, $\mathcal{M}$ is determined by the system itself. However, the coordinate system of $\mathcal{M}$ can be freely chosen. Thus, the time between two successive points on a trajectory of the manifold is not affected if we change the low dimensional coordinate system $\mathcal{M}' \cong \mathcal{M}$. This is because for any mapping:

$$\Delta t_{x_0,x_1} = \int_{\Gamma_{x_0}^{x_1}} \frac{1}{v(x)} dx \qquad (5)$$

In Eq. (5), $\Gamma_{x_0}^{x_1}$ is the trajectory $x(t)$ that connects the two points, and for simplicity we denote $x_i = x(S_i)$. Also, we remind:

$$\frac{\partial S}{\partial t} = \frac{\partial S}{\partial x}\frac{\partial x}{\partial t} \rightarrow v(x) = \frac{\partial S}{\partial t}\frac{\partial x}{\partial S} \qquad (6)$$

Therefore, as long as the cells are on the manifold in the high dimensional space:

$$\Delta t_{x_0,x_1} = \int_{\Gamma_{S_0}^{S_1}} \frac{1}{\frac{\partial S}{\partial t}\frac{\partial x}{\partial S}}\frac{\partial x}{\partial S} dS = \int_{\Gamma_{S_0}^{S_1}} \frac{1}{\dot{S}} dS = \Delta t_{S_0,S_1} \qquad (7)$$

2.3.2. The cell-cycle

In this section, we apply the theory presented above to the cell cycle. We then build a probabilistic formulation to infer the parameters of interest, in particular the cell cycle state dependent phase velocity. To ease the optimization problem, we integrate out the splicing rate parameters.

2.3.2.1. Noise model and dynamics

Assuming multiplicative noise, the noise model for our data can be written as:

$$\begin{aligned} \ln(S_{gc}) &= \ln(s_g(\varphi_c)) + \varepsilon_{gc} \\ \ln(U_{gc}) &= \ln(u_g(\varphi_c)) + \varepsilon_{gc} \end{aligned} \qquad (8)$$

where ε_{gc} comes from a Gaussian probability distribution of mean 0 and standard deviation σ. The corresponding dynamics of the system obeys:

$$\dot{s} = \beta u - \gamma s = \partial_\varphi s(\varphi)\omega(\varphi) \qquad (9)$$

as our coordinate (phase) follows the equation:

$$\dot{\varphi} = \omega(\varphi) \qquad (10)$$

This equation corresponds to Eq. (3) above, with $\omega(\varphi)$ the cell cycle velocity. Given the periodic structure of the data, we write the quantities of interest in Fourier space. To this end, we define:

$$\zeta_f(\varphi) = \delta_{f,0} + \delta_{f-\lfloor\frac{f}{2}\rfloor,1}\cos\left(\left\lceil\frac{f}{2}\right\rceil\theta\right) + \delta_{f-\lfloor\frac{f}{2}\rfloor,0}\sin\left(\left\lceil\frac{f}{2}\right\rceil\theta\right) \quad (11)$$

to be the Fourier base, in which f are the discrete frequencies. We write for the spliced:

$$s_g(\varphi) = e^{\sum_{f=0}^{F}\nu_{gf}\zeta_f(\varphi)} = \sum_{f=0}^{F}\nu'_{gf}\zeta_f(\varphi) \quad (12)$$

and for the unspliced:

$$u_g(\varphi) = e^{\sum_{f=0}^{F}\eta_{gf}\zeta_f(\varphi)} = \sum_{f=0}^{F}\eta'_{gf}\zeta_f(\varphi) \quad (13)$$

In Eq. (12) and (13), ν, ν', η, η' are the Fourier coefficients parametrized in exponential and linear space. Omitting the number of harmonics F from now on, we can rewrite Eq. (8) as:

$$\begin{aligned}\ln(S_{gc}) &= \sum_f \eta_{gf}\,\zeta_f(\varphi) + \varepsilon_{gc} = \ln\left(\sum_f \nu'_{gf}\zeta_f(\varphi)\right) + \varepsilon_{gc}\\ \ln(U_{gc}) &= \ln\left(\frac{1}{\beta_g}\left(\sum_{ff'}\nu_{gf}\,D_{ff'}\zeta_{f'c}\omega(\varphi_c) + \gamma_g\right)s_g(\varphi_c)\right) + \varepsilon_{gc}\end{aligned} \quad (14)$$

with $D_{ff'}$ the Fourier differential operator defined in Eq. (11).

2.3.2.2. A possible problem of scale

The equations written above would be correct if we were able to measure the spliced and unspliced mRNAs in a cell in absolute scale. Not only is this not true because of the low sampling efficiency of modern scRNA-seq techniques, but this sampling efficiency is itself not the same for spliced and unspliced counts. Therefore, we suppose that we measure only a fraction F_s of the spliced counts $\tilde{s}_g$, and F_u of the unspliced counts $\tilde{u}_g$:

$$\begin{aligned}S_{gc} &= F_s\tilde{s}_g(\varphi_c) + \varepsilon_{gc} = s_g(\varphi_c) + \varepsilon_{gc}\\ U_{gc} &= F_u\tilde{u}_g(\varphi_c) + \varepsilon_{gc} = u_g(\varphi_c) + \varepsilon_{gc}\end{aligned} \quad (15)$$

To keep consistency with Eq. (8) written above, we set:

$$\dot{\tilde{s}} = \beta\tilde{u} - \gamma\tilde{s} \rightarrow \dot{s}\frac{1}{F_s} = \beta u\frac{1}{F_u} - \gamma s\frac{1}{F_s} \rightarrow \dot{s} = \beta u\frac{F_s}{F_u} - \gamma s \quad (16)$$

True quantities can be recovered using the corrected value for β, i.e. $\beta \frac{F_S}{F_u}$. In practice, β can be understood as an inverse time scale which will relate quantitatively the velocity we infer with the real velocity. As the global scale is free, this substitution has no impact.

2.3.2.3. Probability distributions and objective function

The joint probability for the spliced and unspliced count measures is $\mathcal{P}(U, S|\{\varphi\}, \{\nu\}, \omega(\varphi), \{\beta\}, \{\gamma\})$. However, S depends only on $(\{\varphi\}, \{\nu\})$, while U depends on $(\{\varphi\}, \{\nu\}, \omega(\varphi), \{\beta\}, \{\gamma\})$. The joint probability can thus be rewritten as:

$$\begin{aligned}&\mathcal{P}(U, S|\{\varphi\}, \{\nu\}, \omega(\varphi), \{\beta\}, \{\gamma\}) = \\ &\mathcal{P}(U|\{\varphi\}, \{\nu\}, \omega(\varphi), \{\beta\}, \{\gamma\})\mathcal{P}(S|\{\varphi\}, \{\nu\})\end{aligned} \qquad (17)$$

Since these probabilities factorize, $\mathcal{P}(S|\{\varphi\}, \{\nu\})$ can be used to infer first the $\{\varphi\}$ and consequently $\{\nu\}$. The remaining $\mathcal{P}(U|\{\varphi\}, \{\nu\}, \omega(\varphi), \{\beta\}, \{\gamma\})$ can be integrated over β to reduce the number of parameters to optimize. To this end, we use a log-normal prior of mean $\bar{\beta}$ and standard deviation τ. Combining these last steps yields (Supplementary Information):

$$\mathcal{P}\left(U\middle|\{\varphi\}, \{\nu\}, \omega(\varphi), \{\gamma\}, \bar{\beta}, \tau\right) \sim \prod_g e^{-\frac{1}{2\sigma^2}\Sigma_c\left(\ln\left(\frac{\bar{u}_{gc}}{U_{gc}}\right)-\langle\ln(\tilde{\beta}_g)\rangle\right)^2} \qquad (18)$$

with

$$\langle\ln(\tilde{\beta}_g)\rangle = \frac{\frac{1}{\sigma^2}\Sigma_c \ln\left(\frac{\bar{u}_{gc}}{U_{gc}}\right) + \frac{1}{\tau^2}\ln(\bar{\beta})}{\left(\frac{|C|}{\sigma^2} + \frac{1}{\tau^2}\right)} = \frac{\Sigma_c \ln\left(\frac{\bar{u}_{gc}}{U_{gc}}\right)}{\sigma^2\left(\frac{|C|}{\sigma^2} + \frac{1}{\tau^2}\right)} \qquad (19)$$

$\mathcal{P}(U|\{\varphi\}, \{\nu\}, \omega(\varphi), \{\gamma\}, \bar{\beta}, \tau)$ is a conditional likelihood function of the two variables of interest, the angular speed $\omega(\varphi)$ and the set of degradation rates $\{\gamma\}$.

2.4. Results

2.4.1. Simulation

We simulate traces according to the consistency equation (4). We first generate unspliced signals from random Fourier series with N harmonics (Figure 3.2A):

$$u_g(\phi) = \mu_g + \sum_{n=1}^{N} a_{g,n}\cos\phi + b_{g,n}\sin\phi \qquad (20)$$

Using realistic values for the parameters (Supplementary Table 3.1, Supplementary Table 3.2), we integrate the consistency equation to obtain the spliced signal from the unspliced one (Figure 3.2B):

$$s_g(\phi) = \int_0^{\phi} \frac{\beta_g u_g(\theta) - \gamma_g s_g(\theta)}{\omega(\theta)} \mathrm{d}\theta \qquad (21)$$

The integrated signal quickly reaches a stable dynamic (transients decay quickly), in which regular oscillations are observed. By cropping out one full cycle, adding Gaussian noise and random drop-out, one obtains a signal which is close to what would have been obtained from a real scRNA-seq experiment (Figure 3.2C). As expected, running a t-SNE dimensionality reduction on the obtained data yields an obvious circular manifold (Figure 3.2D).

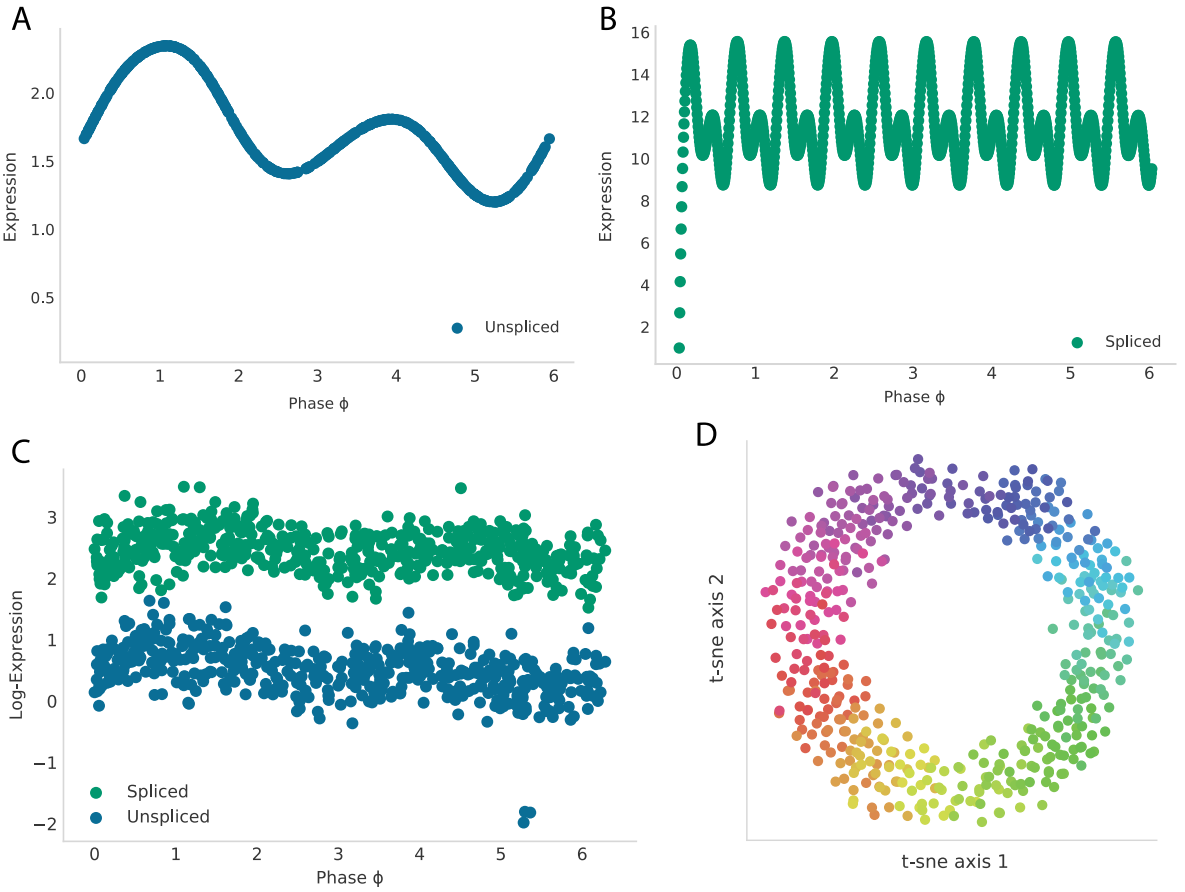


Figure 3.2: Data simulation. (**A**) Unspliced signals are generated from Fourier series with N=4 harmonics. (**B**) Unspliced signals are integrated according to Eq. (21), yielding raw spliced signals exhibiting regular oscillations. (**C**) By cropping out one full cycle, adding noise and drop-out, a realistic spliced signal, as found in scRNA-seq datasets, is obtained. (**D**) The cell-cycle is identifiable from a two-dimensional t-SNE projection of the simulated data.

Inferring the cell-cycle phase using our ad-hoc EM inference method yields excellent results (Figure 3.3). This is somewhat unexpected, as the method was designed to perform optimally with signals having one clearly dominant harmonic, while the simulated dataset uses up to

five harmonics. However, the amplitude of this harmonic is generated randomly, the first one usually being the strongest.

Given the cell phases, the optimizer tries to find the best set of parameters $\{\omega_i, \gamma_g\}$ using the L-BFGS-B method, the corresponding $\{\beta_g\}$ being estimated. To this end, a compromise is made between the best possible fits for the unspliced signal (Supplementary Figure 3.1, Supplementary Figure 3.2), and the prior distribution for β (Methods). An excellent correlation is found between the simulated and estimated values of both $\{\gamma_g\}$ and $\{\beta_g\}$ (Figure 3.3B-C), although the scale is different. This is expected since the formulated model (Eq. (**14**)) is scale-free. The mean $\bar{\beta}$ of the prior for $\{\beta_g\}$ should normally fix the scale, but the optimizer is not powerful enough to find the corresponding minimum. Enforcing the prior confidence can force the absolute scale, but at the expense of the estimated $\{\beta_g\}$ distribution, which becomes too tight to accurately represent the data. In practice, there's no simple way to estimate the (absolute value of the) $\{\beta_g\}$ or $\{\gamma_g\}$ from an experimental dataset, even tough transcript half-live have been measured either for specific genes or in certain cell types, genome-wide. Consequently, we provide the inferred cell-cycle speed in units of the geometric average estimated degradation rate (Figure 3.3D).

Note that, to ensure consistency, the simulation is done by integrating the unspliced profile over the phase, while in the model the unspliced counts are computed as a combination of the derivative of the spliced and of the spliced signal itself.

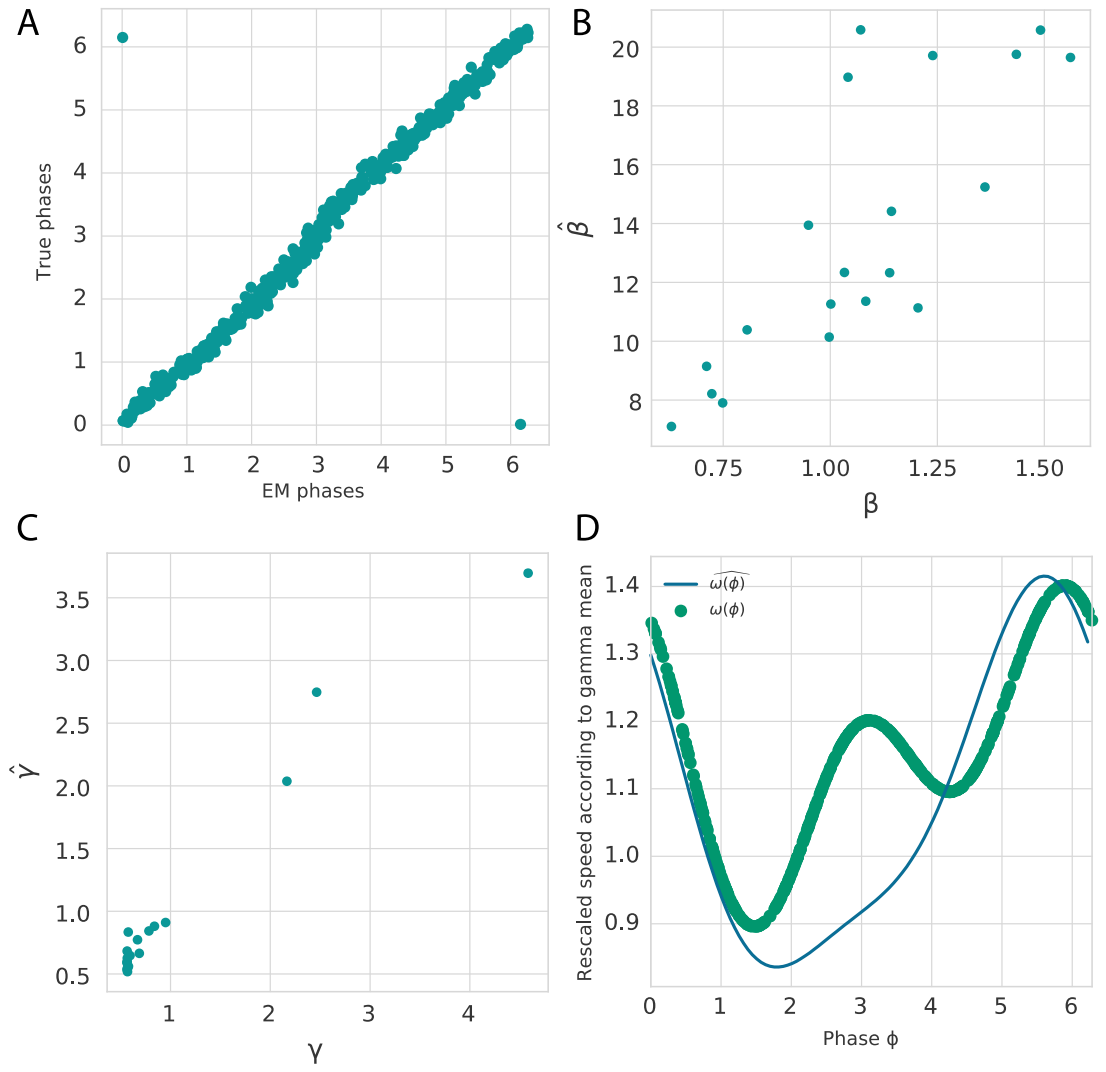


Figure 3.3: Inference results. (**A**) Scatter plot of the simulated *vs* inferred phases (in radians) shows a quasi-perfect correlation (R^2>0.98 in 10 different trials). (**B**) Scatter plot of the estimated *vs* simulated splicing rates (theoretically in h^{-1}, but the scale is free) shows an overall good correlation (R^2>0.7 in 10 separate trials). (**C**) Scatter plot of the estimated *vs* simulated degradation rates (theoretically in h^{-1}, but the scale is free) shows an overall good correlation (R^2>0.6 in 10 different trials). (**D**) Inferred cell-cycle velocity function (blue curve) superimposed with simulated cell-cycle velocity function (green dots) reveals a good but not perfect agreement between the two.

2.4.2. Experimental datasets

In order to characterize the RNA velocity of single cells during the cell cycle, four scRNA-seq datasets were curated for analysis. These datasets differ in their species, tissue source, and sample context. Two datasets are of developing tissues in mice, one of retinal neuron progenitors[271] and one of pancreatic ductal progenitors[283]. On the other hand, two collected datasets are from human adult fibroblasts in untreated and lipid-reducing treatment conditions (unpublished, collected by Irina Khven in the La Manno and D'Angelo

labs). All samples were sequenced using 10X Genomics technology and have been previously assessed in their respective.

For initial preprocessing, unnormalized count matrices were downloaded and using the published UMAP and cell type annotations, any non-cycling cells were removed. Spliced and unspliced counts were obtained either from prior publications or using velocyto[251]. Cells were filtered based on the requirement that the number of unspliced UMIs was greater than the 80th percentile in the data, as to remove cells with too low unspliced reads. Approximate cell cycle phases were assigned to the cells using the signature score method in scanpy[284] using a platform of well-characterized cell cycle marker genes and a previously described algorithm[237], [285]. The final datasets comprised of the following: 968 cells of retina neuron progenitors (S: 406, G2M: 338, G1: 224), 781 cells of pancreatic ductal progenitors (S: 264, G2M: 213, G1: 304), 1,212 untreated fibroblast cells (S: 400, G2M: 449, G1: 363), and 1,397 treated fibroblast cells (S: 464, G2M: 404, G1: 529).

Dimensionality reduction and RNA velocity analysis was then conducted for the datasets using velocyto, resulting in UMAP representations on which movement of cells in gene expression space along a circular trajectory was clearly apparent (Figure 3.4A). Marker genes corresponding to the cell cycle phase boundaries further confirm the presence of a marked cell cycle (Figure 3.4B). Furthermore, trends in cell density, the number of raw reads per cell, and the number of expressed genes were observed in a cell cycle phase-specific manner. There is an increase in the number of UMIs and expressed genes during the cell cycle, with G2M cells having the most reads, followed by a drop in UMI count in freshly-divided G1 cells (Figure 3.4C). This is consistent with the fact that a cell's RNA molecules are split in half upon formation of two daughter cells. Furthermore, there are few cells observed in G2M (Figure 3.4A), and since the scRNA-seq datasets most likely consist of random samples of cells from an unsynchronized population, this suggests that the duration spent in mitosis is much shorter than that in G1 or the growth phase.

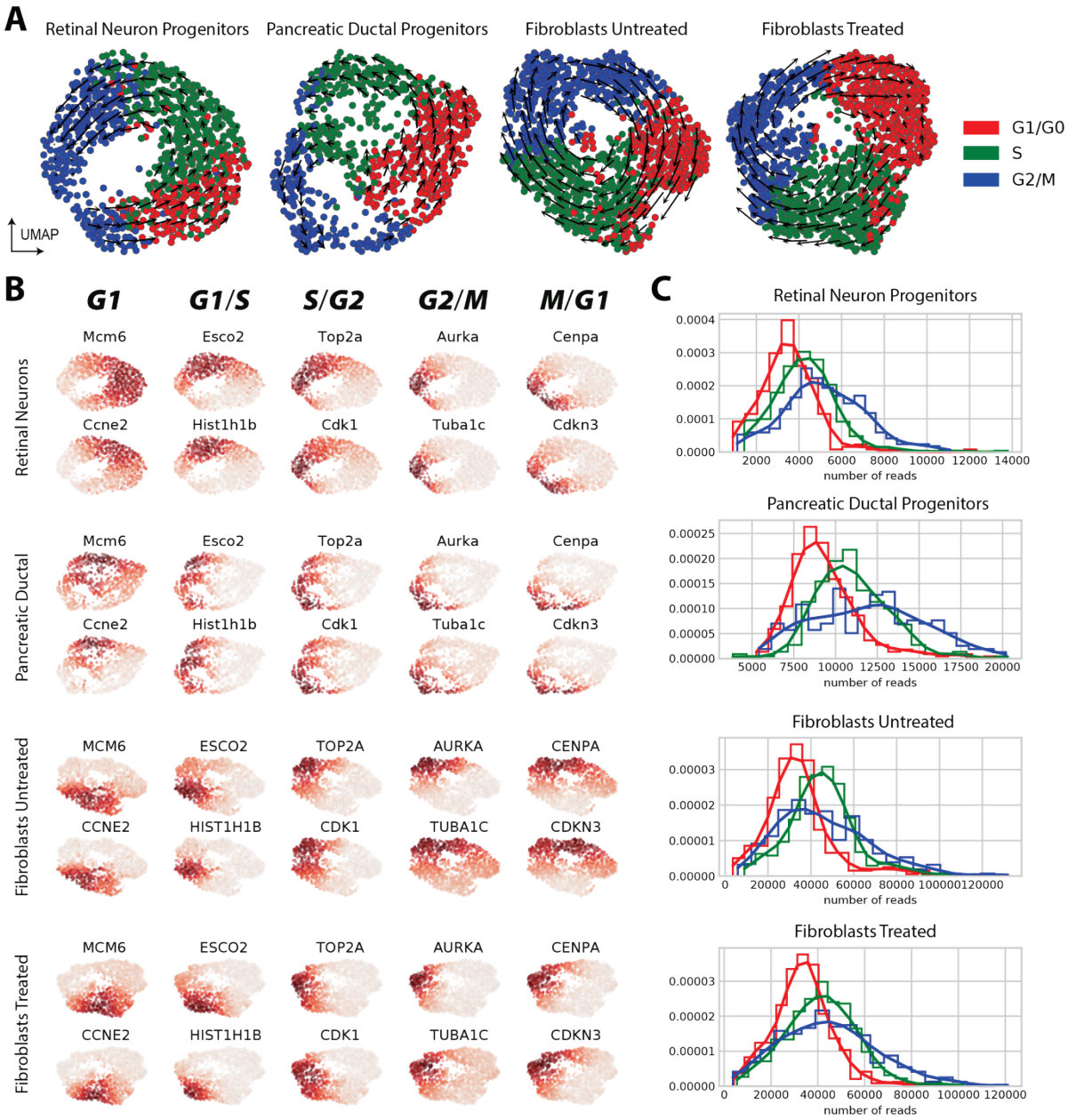


Figure 3.4: Exploratory analysis of the four datasets used with our inference method, including retinal neuron progenitors, pancreatic ductal progenitors, untreated and treated fibroblasts. (**A**) Representations of RNA velocity analysis on UMAP projections reveal clear circular trajectories of the cells in gene expression space (**B**) UMAP representation (same projection as in (A)) of the expression of pre-selected cell-cycle marker genes further confirm the presence of a strong cycling behavior. (**C**) Total number of raw reads quantified on a per cell-cycle phase basis (same colour code as in (A)) reveals a clear dependence of the total gene expression on the cell-cycle phase, with few disparities among the selected datasets.

We then computed the individual gene phases for each dataset. To this end, we extracted a list of 252 cell-cycles genes from the article by *Mizuno et al.*[286]. This comprises 31 genes maximally expressed in growth phase 1 (G1), 20 in-between G1 and synthesis (S) phases, 72 in S phase, 7 in between S and growth phase 2 (G2), 52 in G2 phase, 65 at mitosis (M), and

22 in-between M and G1 phases. The corresponding gene profiles were obtained by running the phase inference method developped by *Talamanca et al.* (Naef-lab, unpublished). We filtered out the profiles that did not oscillate (less than two folds between 5^{th} and 95^{th} percentile), or that showed more than 15% of zeros. In addition, we looked at the cross-correlation between the spliced and unspliced signals, and only kept genes for which the angle corresponding to maximumal cross-correlation was superior to 0.1 radians, as genes without sufficient shifts are not informative for cell cycle dynamics occuring on the time scale of hours [234]. After this selection, between 60 and 120 genes were left per dataset, whose expression was ensured to cover the whole cycle (at least 5 genes per cell-cycle phase, among G1, S, G2, M).

Running our inference methods on the obtained dataset yields quite unexpected results (Figure 3.5), as the estimates for the degradation rate seemed to systematically follow a bimodal distribution, in logarithmic scale. This could be due to biological considerations, e.g. active or passive degradation, but, perhaps more likely, this reflects an inference artefact. Reassuringly, the relative value of the degradation and splicing rate seem more or less conserved across datasets, with mRNAs being degraded about a hundred times more slowly than they are spliced (recall that only the relative quantities are informative since the model is scale free). Finally, the inferred cell-cycle velocities, rescaled according the geometric mean of the corresponding inferred degradation rates, are all of the same order of magnitude, and overall, the speed does not seem to significantly change depending on the phase of the cycle. The individual profiles look qualitatively different, but the differences could be probably attributed to noise, as they didn't seem to follow a clear pattern, and were ony of the order of a few percents relative to the mean. Assuming that the average degradation rate is of the order of 1h^{-1}[287], and the average cell-cycle period is of the order of 20h, this would yield $\frac{\omega}{\gamma} \simeq 0.3$. Yet, depending on the dataset, we find a speed which is about 30% less important. This could mean that the cell under study are slowly cycling, or, alternatively, that an average degradation rate of the order of more than 1h^{-1} is more realistic. If not, this discrepency may point out an inadequation between the model and the dataset, or possibly an optimization problem.

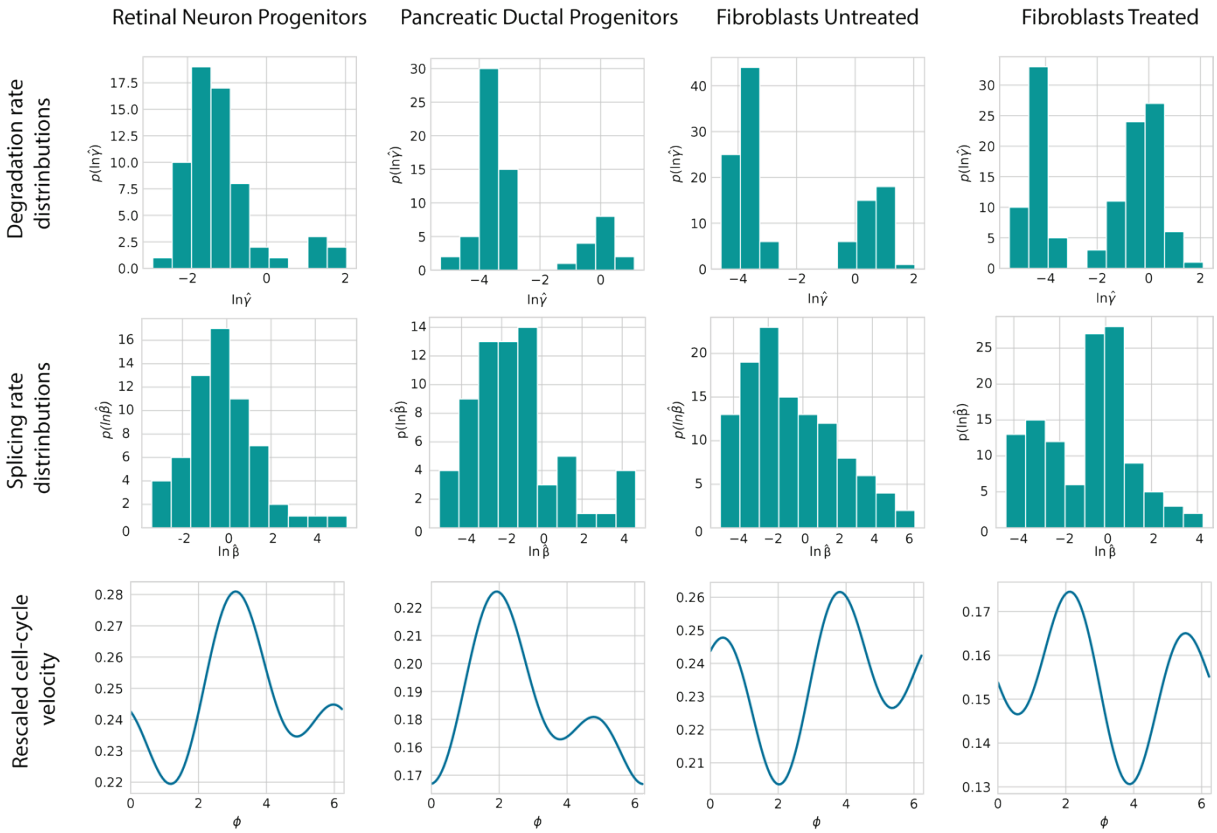


Figure 3.5: Inference results on exprimental datasets from scRNA-seq experiments. (**Top**) Inferred degradation rate (γ) distribution, in logarithmic scale. (**Middle**) Splicing rate (β) estimates distributions, in logarithmic scale. (**Bottom**) Inferred cell-cycle velocity rescaled according to the geometric mean of the inferred degradation rate distribution.

2.5. Discussion

One goal in single-cell transcriptomics is to obtain realistic, low-dimensional, vector field representations of the underlying dynamical process governing single-cells trajectories in expression space. In practice, this dynamic representation is hard to obtain since the data is almost always static, and predictions are made assuming system ergodicity. RNA velocity has revolutionized this approach, but at the cost of strong hypotheses and extensive data wrangling.

Here, we decided to drastically change the approach by using a highly constrained bottom-up method, in which we preliminarily assume the existence of a parametrized manifold, and, *a posteriori*, infer how cells behave on this manifold.

We applied our method to one of the best characterized cell processes, paramount to evolution, and living on a basic circle-like manifold: the cell-cycle. Although our method works very well in simulations, enabling to infer an accurate phase-dependent cell-cycle speed, its efficacy remains somewhat restricted with real datasets, in which the final inferred velocity is lower than expected. We attribute this result to several factors.

First, the inference can optimally work only if the cell phases are adequately inferred. We believe that, although the ad-hoc EM phase inference we've been using performs much better than any other competing method, including trials with personalized UMAP distance function, its inherent imprecision, due to the use of only one harmonic to model the cyclic genes, is still a limiting factor for the quality of our inference.

Then, the optimization process for the degradation rates and cell-cycle speed Fourier coefficients could also be improved using an EM framework. Indeed, the L-BFGS-B optimizer is probably not optimal in a model with so many parameters and constraints. In simulations, it quickly shows limitations when the parameters do not behave closely to what is expected or if too many harmonics are present in the system. Since the model does not involve any unsupervised, black-box steps, an EM framework could probably be implemented and would guarantee a monotonous convergence.

Finally, our method assumes constant splicing and degradation rates, which is a strong hypothesis, somehow questioned during the cell-cycle by recent literature[270]. Unfortunately, embedding the model with time-varying rates seems extremely ambitious, as the total number of parameters to optimize would explode, the rates being presumably gene-specific.

On the other hand, our method can already yield interesting results, as illustrated by the molecular speeds obtained in Figure 3.5. From this plot, our model predicts that the fibrablasts in untreated conditions should cycle faster than those in treated condition, a prediction that could easily be checked experimentally.

In addition, our method offers exciting perspectives, as it could be applied to any manifold, either in replacement or in addition to the cell-cycle. Notably, this includes tree structures, which are most likely the best possible parametrization for development and differentiation processes. Inferring the low-dimensional cell-state could be done with auto-encoders, as already attempted in recent work[288]–[290]. In parallel, embedding the primary manifold, one could study other variables such as circadian or metabolic states.

To our knowledge, this is the first time the cell-cycle velocity is explicitly quantified from static datasets. We expect our work to have a profound impact on the characterization of cell-cycle effects at single-cell resolution, and more generally to greatly aid the analysis of cellular dynamics.

2.6. References

For consistency, references from the article have been placed at the end of the thesis

2.7. Supplementary information

2.7.1. Supplementary methods

2.7.1.1. Gaussian integration

In this section, we briefly present how we integrate the Gaussian distributions used in Eq. (17) of the main text.

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}ax^2+bx+c} = \sqrt{\frac{2\pi}{a}} e^{\frac{b^2}{2a}+c} \qquad (1)$$

$$\int_{-\infty}^{\infty} x\, e^{-\frac{1}{2}ax^2+bx+c} = \frac{b}{a}\sqrt{\frac{2\pi}{a}} e^{\frac{b^2}{2a}+c} \qquad (2)$$

$$\int_{-\infty}^{\infty} x^2\, e^{-\frac{1}{2}ax^2+bx+c} = \frac{(b^2+a)}{a^2}\sqrt{\frac{2\pi}{a}} e^{\frac{b^2}{2a}+c} \qquad (3)$$

$$\int_{-\infty}^{\infty} f(x) e^{-\frac{1}{2}ax^2+bx+c} = \left(e^{\frac{b}{a}\frac{\partial}{\partial x}+\frac{1}{2a}\frac{\partial}{\partial x}\frac{\partial}{\partial x}} f(x)\right)|_0 \sqrt{\frac{2\pi}{a}} e^{\frac{b^2}{2a}\pm c} \qquad (4)$$

$$\begin{aligned} &\int_{-\infty}^{\infty} e^{-\frac{\sum_i (y_i-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\bar{\mu})^2}{2\tau^2}} d\mu = \\ &\int_{-\infty}^{\infty} e^{-\frac{1}{2}\mu^2\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)+\mu\left(\frac{\sum_i y_i}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)-\left(\frac{\sum_i y_i^2}{2\sigma^2}+\frac{\bar{\mu}^2}{2\tau^2}\right)} d\mu = \\ &\sqrt{\frac{2\pi}{\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}}} e^{\frac{\left(\frac{\sum_i y_i}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)^2}{2\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}-\left(\frac{\sum_i y_i^2}{2\sigma^2}+\frac{\bar{\mu}^2}{2\tau^2}\right)} \sim e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\hat{\mu})^2} \end{aligned} \qquad (5)$$

with

$$\hat{\mu} = \frac{1}{\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}\left(\frac{\sum_j y_j}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right) = \frac{\frac{|I|}{\sigma^2}\langle y\rangle + \frac{1}{\tau^2}\bar{\mu}}{\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}} \qquad (6)$$

To normalize a Gaussian probability, the following relationship must hold: $c = -b^2/2a$. Consequently, the only relevant terms are quadratic and linear. Summing up:

$$\mathcal{P}(x) \sim e^{-\frac{1}{2}ax^2+bx+c} \rightarrow \mathcal{P}(x) = \sqrt{\frac{a}{2\pi}} e^{-\frac{1}{2}a\left(x-\frac{b}{a}\right)^2} \qquad (7)$$

Ideally, the exponent of Eq. (5) should also be rewritten more intuitively:

$$\frac{\left(\frac{\sum_i y_i}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)^2}{2\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}-\left(\frac{\sum_i y_i^2}{2\sigma^2}+\frac{\bar{\mu}^2}{2\tau^2}\right)=\frac{a}{2}\sum_i (y_i-\bar{y})^2 \qquad (8)$$

for some a and $\bar{y}$. In particular, we would like to have $\bar{y}=\hat{\mu}$ where $\hat{\mu}=f(\sum_i y_i)$. This yield:

$$\begin{aligned}
&\frac{\left(\frac{\sum_i y_i}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)^2}{2\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}-\left(\frac{\sum_i y_i^2}{2\sigma^2}+\frac{\bar{\mu}^2}{2\tau^2}\right)= \\
&\frac{1}{2\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}\left(\frac{\sum_i y_i}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)\left(\frac{\sum_i y_i}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)-\left(\frac{\sum_i y_i^2}{2\sigma^2}+\frac{\bar{\mu}^2}{2\tau^2}\right)= \\
&-\sum_i\left(\frac{1}{2\sigma^2}y_i^2+\frac{1}{\sigma^2}y_i\frac{1}{\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}\left(\frac{\sum_j y_j}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)\right)+c\rightarrow \qquad (9)\\
&-\frac{1}{2\sigma^2}\sum_i\left(y_i-\frac{1}{\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}\left(\frac{\sum_j y_j}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)\right)^2= \\
&-\frac{1}{2\sigma^2}\sum_i (y_i-\hat{\mu})^2
\end{aligned}$$

Where:

$$\hat{\mu}=\frac{1}{\left(\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}\right)}\left(\frac{\sum_j y_j}{\sigma^2}+\frac{\bar{\mu}}{\tau^2}\right)=\frac{\frac{|I|}{\sigma^2}\langle y\rangle+\frac{1}{\tau^2}\bar{\mu}}{\frac{|I|}{\sigma^2}+\frac{1}{\tau^2}} \qquad (10)$$

which is nothing more than a weighted mean of the data and prior contributions, in which the weights are the variances scaled by the observations. Concisely:

$$\int_{-\infty}^{\infty} e^{-\frac{\sum_i (y_i-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\bar{\mu})^2}{2\tau^2}} d\mu=\sqrt{\frac{2\pi}{\frac{\| I|}{\sigma^2}+\frac{1}{\tau^2}}} e^{-\frac{1}{2\sigma^2}\sum_i (y-\hat{\mu})^2} \qquad (11)$$

2.7.1.2. Application to the model

Given the periodic structure of our system, we can write differentiation as a linear operator in Fourier space. This leads to:

$$\partial_\varphi s_g(\varphi)|_{\varphi=\varphi_c}=\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\omega(\varphi_c)s_g(\varphi_c) \qquad (12)$$

With $D_{ff\prime}$ defined as:

$$D_{ff\prime} = \left(\delta_{\left\lfloor\frac{f}{2}\right\rfloor,\left\lfloor\frac{f'}{2}\right\rfloor} - \delta_{f,f'}\right)\left\lceil\frac{f}{2}\right\rceil(-1)^f \qquad (13)$$

Inserting in Eq. (4) of the main text, we have:

$$\ln(U_{gc}) = \ln\left(\frac{1}{\beta_g}\left(\sum_{ff\prime}\nu_{gf}\,D_{ff\prime}\zeta_{f\prime c}\omega(\varphi_c) + \gamma_g\right)s_g(\varphi_c)\right) + \varepsilon_{gc} \qquad (14)$$

Consequently:

$$\begin{aligned}&\mathcal{P}(U|\{\varphi\},\{\nu\},\omega(\varphi_c),\{\beta\},\{\gamma\}) \sim \\ &\exp\left(-\frac{1}{2\sigma^2}\sum_{gc}\left(\ln(U_{gc}) - \ln\left(\frac{1}{\beta_g}\left(\sum_{ff\prime}\nu_{gf}\,D_{ff\prime}\zeta_{f\prime c}\omega(\varphi_c) + \gamma_g\right)s_g(\varphi_c)\right)\right)^2\right) \sim \\ &\exp\left(-\frac{1}{2\sigma^2}\sum_{gc}\left(\ln(U_{gc}) + \ln(\tilde{\beta}_g) - \ln\left(\left(\sum_{ff\prime}\nu_{gf}\,D_{ff\prime}\zeta_{f\prime c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right)s_g(\varphi_c)\right)\right)^2\right)\end{aligned} \qquad (15)$$

In Eq. (16) of the main text, we have introduced the unitless notation: $\tilde{\beta}_g = \beta_g/\bar{\beta}$, $\tilde{\gamma}_g = \gamma_g/\bar{\beta}$, $\tilde{\omega}(\varphi) = \omega(\varphi)/\bar{\beta}$. $\bar{\beta}$ is the mean of the prior probability distribution. Ideally, we would like to rewrite the probability of the unspliced as:

$$\mathcal{P}(U|\{\varphi\},\{\nu\},\omega(\varphi),\{\beta\},\{\gamma\}) \sim \exp\left(\sum_g\left(-\frac{1}{2}a_g\ln(\tilde{\beta}_g)^2 + b_g\ln(\tilde{\beta}_g) + c_g\right)\right) \qquad (16)$$

We need to calculate the gene-specific coefficients that appear in the integral: (a, b, c).

$$\begin{aligned}&\sum_{gc}\left(\ln(U_{gc}) + \ln(\tilde{\beta}_g) - \ln\left(\left(\sum_{ff\prime}\nu_{gf}\,D_{ff\prime}\zeta_{f\prime c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right)s_g(\varphi_c)\right)\right)^2 = \\ &\sum_{gc}\left(\ln(\tilde{\beta}_g) - \ln\left(\frac{\left(\sum_{ff\prime}\nu_{gf}\,D_{ff\prime}\zeta_{f\prime c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right)s_g(\varphi_c)}{U_{gc}}\right)\right)^2 = \\ &|C|\ln(\tilde{\beta}_g)^2 - \sum_{gc}2\ln(\tilde{\beta}_g)\ln\left(\frac{\left(\sum_{ff\prime}\nu_{gf}\,D_{ff\prime}\zeta_{f\prime c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right)s_g(\varphi_c)}{U_{gc}}\right) + \\ &\sum_{gc}\ln\left(\frac{\left(\sum_{ff\prime}\nu_{gf}\,D_{ff\prime}\zeta_{f\prime c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right)s_g(\varphi_c)}{U_{gc}}\right)^2\end{aligned} \qquad (17)$$

This directly means:

$$\begin{aligned} a_g &= \frac{|C|}{\sigma^2} \\ b_g &= \frac{1}{\sigma^2}\sum_c \ln\left(\frac{\left(\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right)s_g(\varphi_c)}{U_{gc}}\right) \\ c_g &= \frac{-1}{2\sigma^2}\sum_c \ln\left(\frac{\left(\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right)s_g(\varphi_c)}{U_{gc}}\right)^2 \end{aligned} \qquad (18)$$

For now, we can neglect c_g, as it is only a normalization constant. To compute

$$\int \mathcal{P}(U|\{\varphi\},\{\nu\},\omega(\varphi),\{\beta\},\{\gamma\})\mathcal{P}(\{\beta\}|\{\varphi\},\{\nu\},\omega(\varphi),\{\gamma\})d\beta \qquad (19)$$

we need a (conditioned) prior on β. For ease of analytics, we use a log-normal distribution (instead of, ideally, a gamma or beta one). The mean of this prior should not be very informative as it sets a time scale, and the system is time-invariant. For simplicity, we set it to zero. However, the corresponding spread τ is important. The distribution is:

$$\mathcal{P}(\{\beta\}|\{\varphi\},\{\nu\},\omega(\varphi),\{\gamma\}) \sim \prod_g \frac{1}{\tilde{\beta}_g} e^{-\frac{1}{2\tau^2}\ln(\tilde{\beta}_g)^2} \sim \prod_g \frac{1}{\beta_g} e^{-\frac{1}{2\tau^2}\ln(\beta_g/\bar{\beta})^2} \qquad (20)$$

We can now explicitly write the integral:

$$\begin{aligned} &\int \mathcal{P}(U|\{\varphi\},\{\nu\},\omega(\varphi),\{\beta\},\{\gamma\})\mathcal{P}(\{\beta\}|\{\varphi\},\{\nu\},\omega(\varphi),\{\gamma\})d\beta \sim \\ &\prod_g \int e^{\left(-a_g\ln(\tilde{\beta}_g)^2 + b_g\ln(\tilde{\beta}_g) + c_g\right)} e^{-\frac{1}{2\tau^2}\ln(\tilde{\beta}_g)^2} \frac{1}{\tilde{\beta}_g} d\tilde{\beta}_g \sim \\ &\prod_g \int e^{\left(-a_g\ln(\tilde{\beta}_g)^2 + b_g\ln(\tilde{\beta}_g)\right)} e^{-\frac{1}{2\tau^2}\ln(\tilde{\beta}_g)^2} d\ln(\tilde{\beta}_g) = \\ &\prod_g \int e^{\left(-\frac{1}{2}(a_g+\frac{1}{\tau^2})\ln(\tilde{\beta}_g)^2 + b_g\ln(\tilde{\beta}_g)\right)} d\ln(\tilde{\beta}_g) \sim \prod_g e^{\frac{(b_g)^2}{\frac{1}{2}\left(a_g+\frac{1}{\tau^2}\right)}} \end{aligned} \qquad (21)$$

Integrate out $\tilde{\gamma}$ requires some approximations. Especially, $e^{\sum_g c_g}$ must be considered. We have:

$$\mathcal{P}(U|\{\varphi\},\{\nu\},\omega(\varphi),\{\gamma\}\bar{\beta},\tau) \sim \prod_g e^{\frac{(b_g)^2}{\frac{1}{2}\left(a_g+\frac{1}{\tau^2}\right)}+c_g} \qquad (22)$$

To simplify further calculations, let's introduce some new variables:

$$\bar{u}_g = u_g(\beta_g = \bar{\beta}) = \frac{\beta_g}{\bar{\beta}} u_g = \left(\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c) + \tilde{\gamma}_g\right) s_g(\varphi_c) \qquad (23)$$

Using the equalities proved in Eq. (9), we obtain:

$$\mathcal{P}(U|\{\varphi\},\{\nu\},\omega(\varphi),\{\gamma\},\bar{\beta},\tau) \sim \prod_g e^{-\frac{1}{2\sigma^2}\sum_c\left(\ln\left(\frac{\bar{u}_{gc}}{U_{gc}}\right)-\langle\ln(\tilde{\beta}_g)\rangle\right)^2} \qquad (24)$$

with

$$\langle\ln(\tilde{\beta}_g)\rangle = \frac{\frac{1}{\sigma^2}\sum_c \ln\left(\frac{\bar{u}_{gc}}{U_{gc}}\right)+\frac{1}{\tau^2}\ln(\bar{\beta})}{\left(\frac{|C|}{\sigma^2}+\frac{1}{\tau^2}\right)} = \frac{\sum_c \ln\left(\frac{\bar{u}_{gc}}{U_{gc}}\right)}{\sigma^2\left(\frac{|C|}{\sigma^2}+\frac{1}{\tau^2}\right)} \qquad (25)$$

as $\bar{\beta}=1$.

2.7.1.3. The constraints on γ

In simulations, if no care is taken when sampling ν and γ, the following can occur:

$$u_{gc} = \left(\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c)+\tilde{\gamma}_g\right)s_g(\varphi_c) < 0 \qquad (26)$$

Now u_{gc} is necessarily positive. This led us to notice an important relation:

$$\min(\partial_t s) \geq -\gamma s \qquad (27)$$

Therefore, in the model, it must be included that:

$$\begin{aligned} \min_c\left(\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c)s_g\right) &\geq -\gamma_g s_g \\ \rightarrow \quad \min_c\left(\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c)\right) &\geq -\gamma_g \end{aligned} \qquad (28)$$

Which means:

$$\min_c\left(\sum_{ff'} \nu_{gf}\, D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c)+\gamma_g\right) \geq 0 \quad \rightarrow \quad u \geq 0 \quad \text{as} \quad s \geq 0 \qquad (29)$$

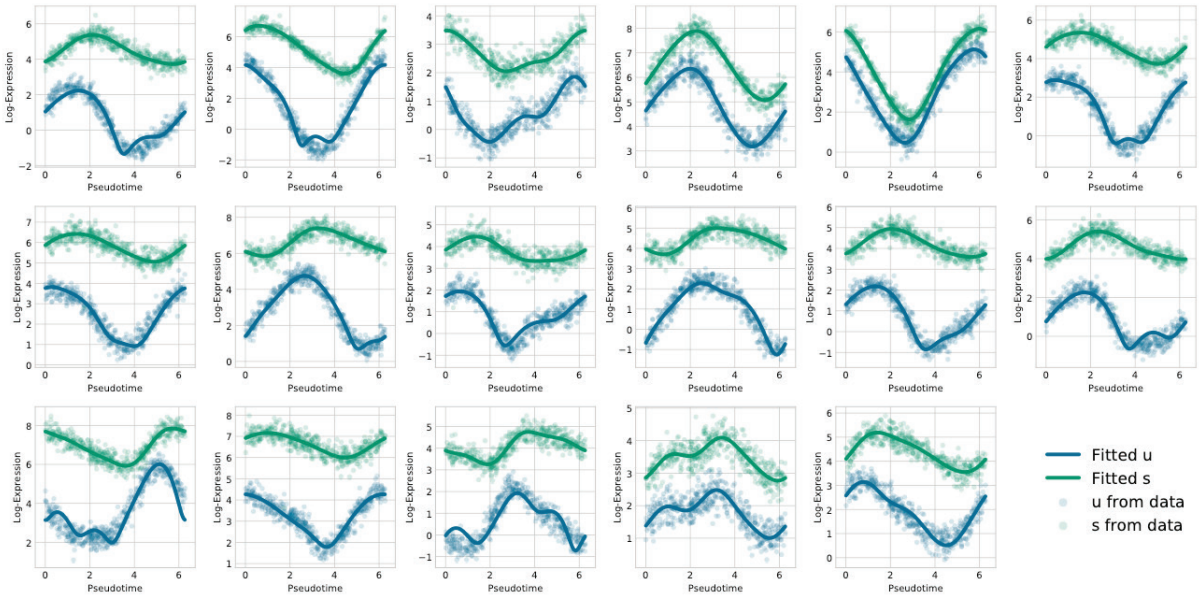
As it seems easier to freely choose ν and then constrain gamma. Accordingly, we calculate:

$$\min_{c}\left(\sum_{ff'}\nu_{gf}\,D_{ff'}\zeta_{f'c}\tilde{\omega}(\varphi_c)\right)\geq$$
$$\min_{\varphi}\left(\sum_{ff'}\nu_{gf}\,D_{ff'}\zeta_{f'}\tilde{\omega}(\varphi)\right)\geq \tag{30}$$
$$-\max_{\varphi}(\omega(\varphi))\sum_{f\ \mathrm{odd}}\left(\frac{f+1}{2}\sqrt{\nu_{g,f}^2+\nu_{g,f+1}^2}\right)$$

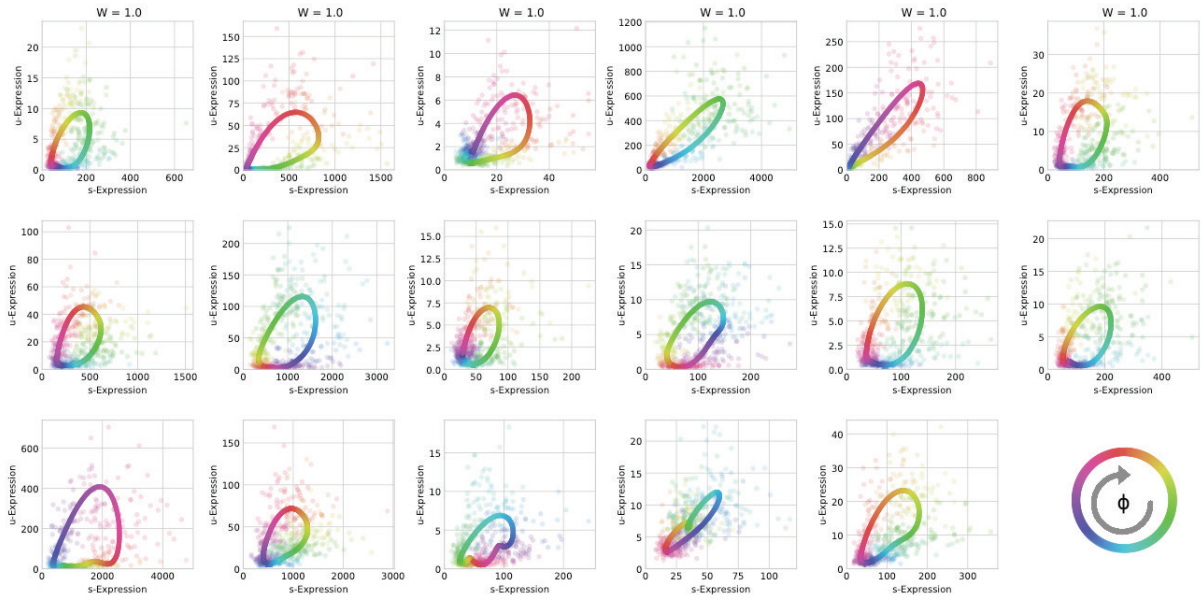
Thus:

$$\gamma_g \geq \max_{\varphi}(\omega(\varphi))\sum_{f\ \mathrm{odd}}\left(\frac{f+1}{2}\sqrt{\nu_{g,f}^2+\nu_{g,f+1}^2}\right) \tag{31}$$

2.7.2. Supplementary figures



Supplementary Figure 3.1: Temporal representation of the fits obtained from the inference. The optimizer first fits the spliced data (green dots) with a Fourier series (green curve). The unspliced data (blue dots) is then fitted (blue curve) by optimizing the degradation rate and the cell-cycle speed such that the corresponding likelihood is maximum.



Supplementary Figure 3.2: Phase-space representation of the fits obtained from the inference. Same as Supplementary Figure 3.1, except that the data (dots) and fits (curves) are represented in the spliced-unspliced space, in which the simulated/inferred phase is represented by a colour (bottom right).

2.7.3. Supplementary Tables

Parameter	Biological range of values (used for simulations)	Value used for rescaling	Rescaled value	Rescaled value (γ units)
Splicing rate β (h^{-1})	$\beta^{-1} \simeq 5mn - 45mn$	$\beta_0 = 4h^{-1}$	$\beta_0 = 1$	$\beta_0 = 8\gamma_0$
Degradation rate γ (h^{-1})	$\gamma^{-1} \simeq 30mn - 5h$	$\gamma_0 = 0.5h^{-1}$	$\gamma_0 = 1/8$	$\gamma_0 = 1\gamma_0$
Cell-cycle speed ω (h^{-1})	$\omega^{-1}/2\pi \simeq 10h - 30h$	$\omega_0 = 0.30h^{-1}$	$\omega_0 = 0.08$	$\omega_0 = 2/3\gamma_0$

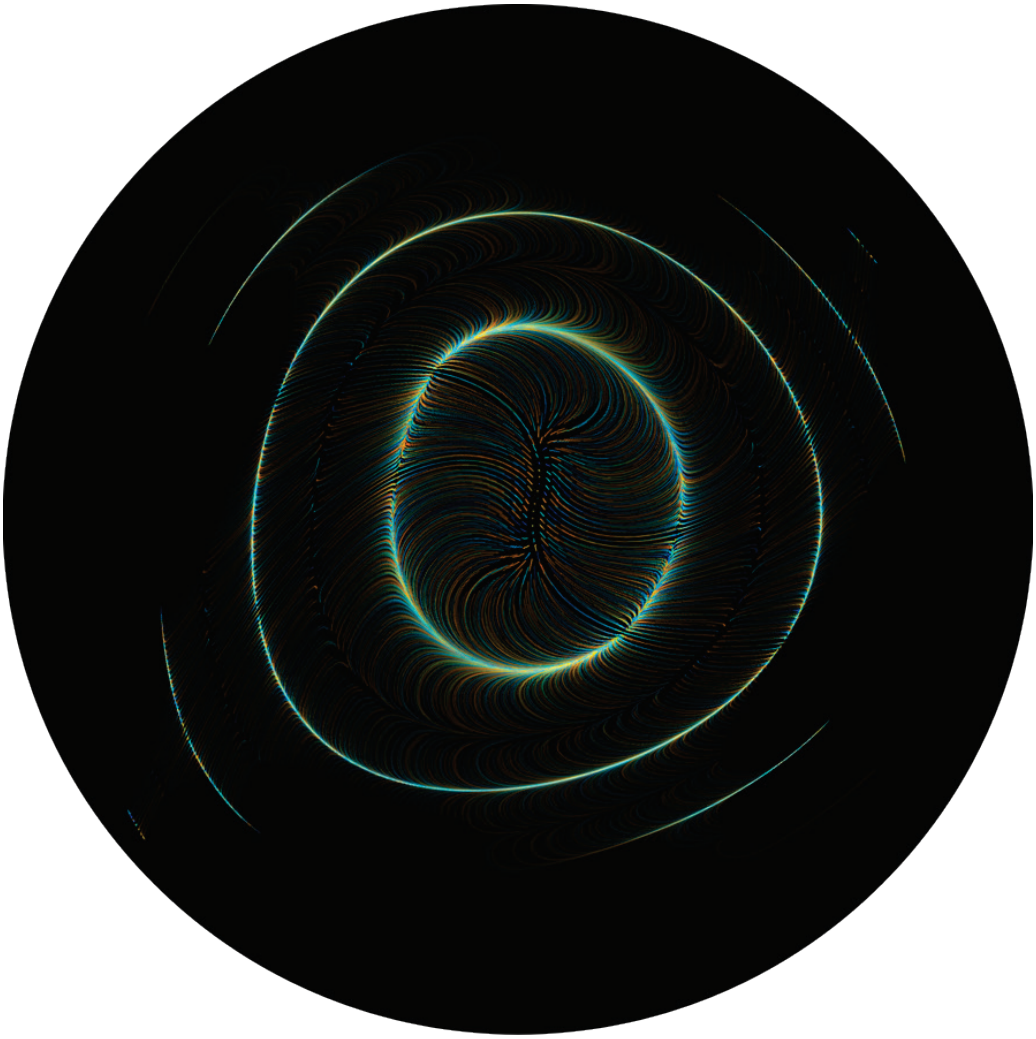
Supplementary Table 3.1: Parameters used for simulation and scaling.

Parameter	Biological range of values (used for simulations)	Value used for inference
Number of harmonics for the spliced signal	$\simeq$4	2
Number of harmonics for ω	$\simeq$4 (?)	2
External noise	$\simeq 0.3 - 0.8$	0.5

Supplementary Table 3.2: Parameters used for simulation and inference.

Discussion and perspectives

This thesis has been essentially concerned with the inference from, along with modelling and analysis of, noisy biological oscillatory data. The Introduction was presented a summarized background of the problems and methods studied in the rest of this document. Chapter 1 to 3 presented three concrete applications of the methods previously introduced, with the study of the interacting cell-cycle and circadian clocks, circadian zonation of gene expression, and cell-cycle speed inference using RNA velocity. The present chapter reflects on the obtained results and points out some potential directions for future research.

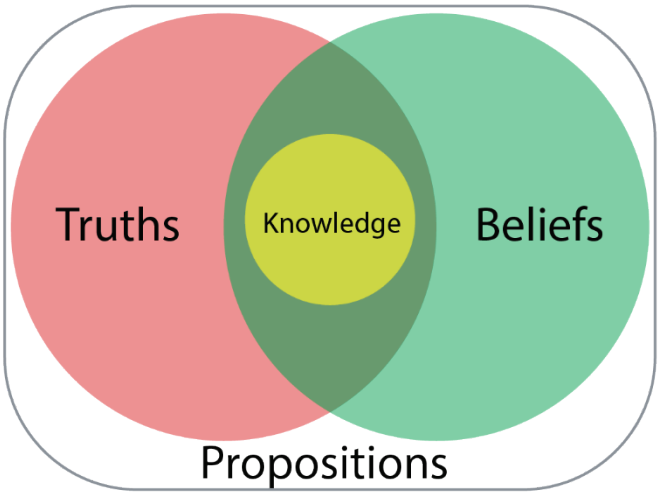


Artwork Figure 5: Artistic representation of an attracting noisy circular vectorfield, with initial conditions uniformly distributed on a grid. Made with Processing, inspired by the work found on *generateme.wordpress.com*.

1. Discussion

1.1. Knowledge acquired

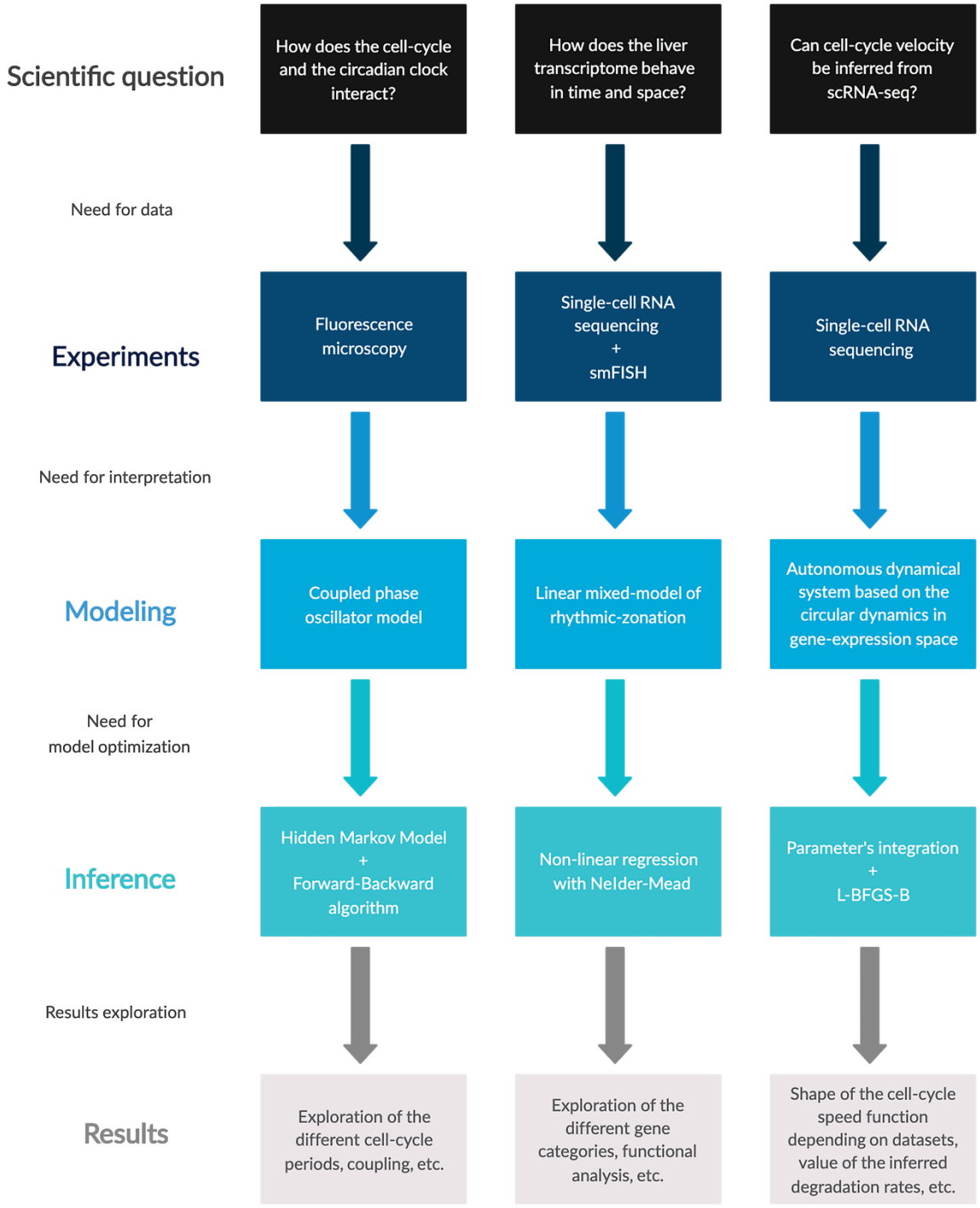
More than two millennia ago, in *Theaetetus*, Plato defined knowledge as a subset of what is both true and believed to be true (Discussion Figure 1). Although this was a small revolution in the nascent field of epistemology, it took a long time to develop a systematic methodology to decipher the true value of a belief. Between the XVth and XXth century, several important concepts, often contradictory, were theorized to this end: scepticism, rationalism, inductivism, and finally hypothetico-deductivism[291]. Although all of these are still part of the scientific arsenal, the main innovation was probably the creation of Bayesian epistemology, in the XVIIIth century. What was initially considered a field of philosophy has become, in the XXth century, a major scientific tool, with the use of probability laws as coherent constraints on rational degrees of confidence, as well with the introduction of probabilistic inference[292].



Discussion Figure 1: Euler's diagram of the relationship between truths, beliefs and knowledge, in platonic epistemology.

The work presented in this thesis illustrates a concrete application of these epistemological principles (Discussion Figure 2). We started from experimental biological data: fluorescence traces in Chapter 1, followed by scRNA-seq in Chapter 2 and 3. We then developed different models, whose structures themselves constitute a prior belief about certain statistical properties of the data: models of coupled phase oscillators in Chapter 1; mixed-effects models using time-dependent polynomes to represent rhythmic gene expression in Chapter 2; and autonomous dynamical systems making use of a tangent space to the RNA velocity manifold in Chapter 3. Finally, we inferred the model parameters with appropriate methods: HMM along with the forward-backward algorithm in Chapter 1, non-linear regression with Nelder-

Mead in Chapter 2 and L-BFGS-B in Chapter 3. We were then able to explore model parameters to make some predictions on the systems under study: phase-lockings in Chapter 1, expected zonation profiles in Chapter 2 and expected cell-cycle velocity in Chapter 3.



Discussion Figure 2: Diagram of the methodological approach followed to answer the scientific questions asked in this thesis.

None of the approaches presented above is a black-box, as the models used have the property of being interpretable in terms of explicit dynamical or physical models. That is, one can

understand, from a mathematical point of view, what happens in these systems. Still, to which extent are the results we obtained informative? It is often thought in epistemology that no model will ever truly capture reality, as the (ontological) nature of the world can only be captured partially, through measurement devices[293]. Yet, mathematical models can yield results of astonishing precision, predicting phenomena millions of years in advance (e.g. in astronomy), especially in noiseless systems. Similarly, in biology, the application of simple principles (e.g. Evolution theory) has led to significant discoveries, particularly in medicine (e.g. cancer initiation, progression, treatment, and resistance). Therefore, a contradiction seemingly arises here as, on the one hand, the true nature of the world is unfathomable, and on the other, the scientific method seems able to capture some of its properties.

One possibility to overcome this aporia would be to consider science as only one interpretation of the world, among many others. This interpretation is partial, as it is limited by the set of tools used by scientists but can nevertheless yield astonishing results when it comes to practical application. For instance, physicists normally interpret their data under some well defined assumptions and approximations (e.g. coarse graining, isolated system, etc), and what they learn will be highly conditioned by those. To relate to the content of this thesis, saying that the cell-cycle and the circadian clock interact through a given coupling function does not do justice to what is occurring in the real system, in which billions of atoms interact according to laws whose behaviour is still poorly understood[25]. But it remains a (partial) truth, under the interpretation of dynamical systems, and this truth is undeniably an advance in terms of biological knowledge.

1.2. The contribution of modelling

Most of the research in biology is done in from a purely experimental perspective. A question is asked (e.g. what's the division rate of a given cell-type), an experiment is led (e.g. brightfield microscopy), and the question is almost immediately answered, although some postprocessing of the data may be involved (e.g. image analysis). This type of approach is efficient and straightforward, but has severe limitations; namely, it only works if the question asked concerns simple systems. One could argue that living organisms are in no way simple systems, but, from a reductionist point of view, they can often be reduced to simpler sub-systems.

Now, it happens that some biological questions raise intrinsically complex matters. For instance, if one were to ask how, mechanistically, the cell-cycle is coupled to the circadian

[25] And the extent to which these laws can be put in equations is still an open question.

clock, one wouldn't even know where to look, due to the high number of possibilities. Indeed, many experiments could be made, answering just as many questions. For instance: is NONO involved in the coupling? Is p53 involved in the coupling? Is Cry1 involved in the coupling? But there are thousands of molecules to test, and each experiment should also involve additional parameters such as temperature or pressure. In short, this is not concretely feasible.

Modelling can here help in two different ways. The first one is through the involvement of *universality*, a mathematical observation that some properties are common, independent of the details of the system under study[294]. This is particularly relevant to abstract physical modeling, such as dynamical systems theory, which can explain, with the same type of differential equations, the rotation of the Earth around the sun, and the behaviour of a harmonic oscillator. Similar situations arise in biology, where many different systems can be studied within the same framework, e.g. oscillations in the brain somehow follow the same laws as cyclins in individual cells. Aside from the fact that it is puzzling from a philosophical point of view, universality is also useful to answer questions from partial observations. For instance, in Chapter 1, although we only followed Rev-erbα to track the circadian clock in NIH3T3, we were still able to make general predictions about its interaction with the cell-cycle (namely, how does the dephasing evolves with the cell-cycle period), which we verified thereafter. We didn't know anything about the mechanistic details of the coupling, but we could make predictions nevertheless, without the need for thousands of experiments.

Another way mathematical modelling can help is through the analysis of complex or large datasets. For instance, even if one were to run a thousand experiments to understand all the molecular details of the coupling, one would still have to rebuild the system's dynamics from partial observations. Given how complex genetic networks can be, involving many feedbacks loops themselves under the control of other feedback loops[295], this is hardly feasible in a human lifetime. Modelling could help here by simplifying the system to its most essential version. For instance, dimensionality reduction techniques such as UMAP or SVD could enable identifying which molecules are relevant to the coupling, and get rid of all the others[296].

Nowadays, doing a thousand experiments is not always required anymore. With the emergence of high-throughput methods such as RNA sequencing, or, more recently, single-cell RNA sequencing, complex datasets containing most of the relevant interactions (as well as the non-relevant) are available to everyone: one just needs to find or develop the proper method to extract the relevant information. This can, however, be a hard problem, as illustrated by Chapter 3, in which the theoretical method we developed indeed answers the question, but somehow at the expense of model tractability.

2. Perspectives

Although the studies presented in Chapter 1-3 of this thesis are relatively self-contained, they inherently suggest interesting extensions.

Chapter 1 mainly presented the inference of the coupling function between the cell-cycle and the circadian clock. While the corresponding theoretical model of interacting phase oscillators is deeply explored, this is not the case for the mechanistic factors responsible for the coupling. Since we have posited that the coupling, which mainly occurs at mitosis, is at least partly due to the drop of PER/CRY concentration with the nuclear envelope breakdown, it would be interesting to model how these proteins behave using e.g. reaction-diffusion equations. In parallel, regarding the statistical procedure itself, it could be interesting to turn the powerful phase inference HMM-based method that we had developed into a generic package, so that it becomes available to anyone interested. This is partly what was attempted in the project presented Annexe A, with the implementation of methods handling multi-reporters' systems like FUCCI. This idea was however dropped because of scooping, but a generic library for phase (or coupling) inference is still not available today. Finally, extending the project to spatially coupled oscillator systems (e.g. somite clock, pace-making in neuron networks, organ development) could also be interesting, as they show dynamical properties such as wave propagation and, in most cases, can also be tracked using fluorescent reporters[297].

Chapter 2 presented the reconstruction and spatiotemporal study of gene profiles in the mammalian liver. From an experimental perspective, it could help to use methods examining epigenetic properties of cells such as ATAC-seq and ChIP-seq to determine how the oscillations of Wnt2 (and potentially other genes) are regulated and whether circadian genes bind to wnt-related genes in LECs. In addition, conditional Bmal1 knock-out models targeted at endothelial cells would help to move from descriptive to more mechanistic studies and elucidate causal relationships. From a modelling perspective, moving to a fully Bayesian framework such as Bayesian hierarchical modelling could help to handle the noise[298]. Also, finding a way to remove the spatial correlations inherent to the gene profiles reconstruction (explained in reference [224]) would probably yield more accurate results.

Chapter 3 presented the inference of cell-cycle speed using a new, fully parametric version of RNA velocity. Although the method works relatively well, it still has many drawbacks (the reason for which the study is not a preprint, yet). The main issue is probably with optimization, as the space of parameters is constrained to positive real numbers by the use of logarithm transformation in the model, a constraint which is poorly handled by the optimizer. Since the model is explicitly parametrized, one solution could be to use the Expectation-Optimization algorithm as an alternative optimizer. Unfortunately, this requires quite

complex computations, and would probably take a few months of work, for a gain which could be minimal. Another likely issue with the model lies in the use of constant transcription, splicing, and degradation rates. This is a strong hypothesis, questioned by a recent study[270], and which could significantly impact the value and shape of the inferred cell-cycle speed function. Unfortunately, assuming gene-dependent varying rates, whether for transcription, splicing or degradation (or all) implies involving many new parameters in the model. At the same time, the data is limited, and the optimizer already struggling. This could, however, be at least partly attempted using pulse-chase analysis.

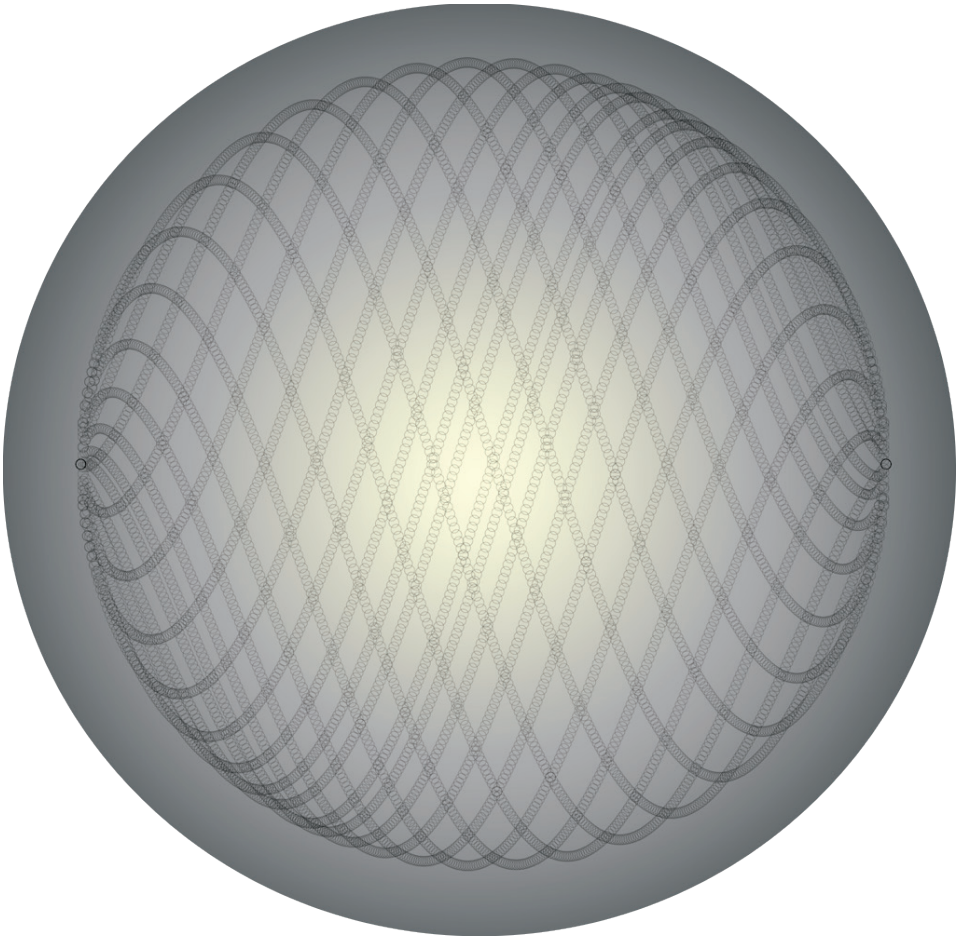
From a more general perspective, the use of modern machine learning methods such as deep neural network will probably revolutionize the field of biological inference, and should be the focus of upcoming studies[299]. This is particularly relevant to image analysis (e.g. fluorescence microscopy[300]), and big data exploration (e.g. omics data[301]). The drawback of these approaches is that they are usually black boxes, preventing interpretation of the underlying model. However, this is progressively changing with the development of powerful regularization techniques enabling to keep only the most essential parameters, and the development of neural network exploration techniques, allowing to understand better the role of the different layers and layer components[302].

Annexe A: Technical review of phase inference methods

The chapter that follows is an unfinished comparative study of phase inference methods. This paper was written in my free time and was unfortunately partly scooped in the process[303], [304], the reason for which it will remain unfinished.

Contributions

I designed the study, implemented the code, and made all the figures.



Artwork Figure 6: Artistic representation of progressively dephasing harmonic function, with increasing amplitude. A perfect circle is formed by the final superposition.

1. Introduction: motivation and aims

Oscillators are pervasive in biology, and the study of oscillations is particularly relevant to neuroscience, geology, population dynamics and biochemistry. Yet, these fields are often oblivious of the importance of phase dynamics to understand the behaviour of systems under external coupling or perturbation. For instance, most studies in geology are still using Hilbert transform (or its extensions) for phase inference[305], [306]. The same goes for neuroscience, including recent studies[307]. Now, it is precisely from phase behaviour in response to a given perturbation or coupling that one can learn details about the off steady-state system dynamics.

Although several powerful methods for phase inference have been developed in the last century, there is absolutely no summarizing work or guidelines in the literature about these methods. Worse, most of these methods, like smoothing methods[308] or empirical mode decomposition[309]) are not initially designed for phase inference, and people would not know about them even when extensively looking for a solution.

In this project, we want to introduce and compare the main existing methods used for phase inference in noisy oscillating systems, possibly under perturbations. This includes systems for which several channels of observation are given, and complex coupling link the different oscillators.

We want to keep this review light enough not to confuse the readers with non-technical backgrounds, but complex enough to introduce the crucial concepts needed to understand which method is appropriate depending on the context. We also would like to compare the performances of the different methods under different conditions, summarize their results, and provide the associated code for easy implementation.

2. Study

2.1. Abstract

Oscillating systems are widespread in the biological world, occurring at all spatial and temporal scales. Understanding how these systems react to perturbation requires capturing their instantaneous behaviour, which is provided by their phase. There is, however, no generic phase inference method that works for any signal. Existing methods are often not initially designed for phase inference and, as such, are sometimes hard to implement. We here provide a technical summary of the existing phase inference methods, and briefly compare their performance on simulated data.

2.2. Introduction

2.2.1. Context

Biological oscillators are of prime importance in many different contexts: development, signalling, metabolism, etc. A large portion of the cell physiological processes are thus under the control of biochemical oscillations. This includes circadian rhythms, cell-cycle, but also somitogenesis or neuronal circuits[25], [51]. The steady-state behaviour of biological oscillators is well captured by signal analysis methods such as Fourier transform or wavelet analysis, which decompose the oscillations into various harmonics with different frequencies and amplitude[310]. However, when studying a system under perturbation or transiently evolving, such methods show little interest as they are designed to work on the scale of several stable cycles. Yet, understanding how oscillatory systems react to perturbation requires capturing their instantaneous behaviour, which is provided by their phase and amplitude. Unfortunately, inferring the phase of a partly-randomly evolving oscillatory system is not an easy problem, as a strong perturbation can be understood equally as a shift backwards of forward on the polar cycle. As a consequence, no phase inference method can yield an accurate result for any type of signal.

In this technical review, we provide an exhaustive summary of the phase inference methods described in the literature, with a particular emphasis on smoothing methods. We explain how the methods work and to which context they're the most adapted. Finally, we compare the methods' performances on simulated data, including signals made of several channels. The whole study is implemented in Julia and provided as a collection of Jupyter notebook.

2.2.2. The problem in equations

Assuming that the noise is independent of the phase, as it's the case in most limit-cycle oscillators[1], the most generic dynamical system describing the evolution of a phase oscillator is:

$$\frac{d\theta}{dt} = \omega_\theta + \xi \qquad (i)$$

In this differential equation, θ is the oscillator's phase, ω_θ is the intrinsic frequency of the oscillator and ξ is the intrinsic phase noise. The corresponding signal is described by the following transformation:

$$y_t = s(\theta_t) + \epsilon_t \qquad (ii)$$

Here, y_t corresponds to the data point observed at time t, s is a periodic function of the phase which we'll call *waveform* from now on, and ϵ_t is the technical (observation) noise. Phase inference can be understood as the process of estimating the probability of the phase given the data, $P(\theta_t|y_1, y_2, \ldots, y_T)$. Alternatively, in a simpler setting, it can correspond to getting a correct pointwise estimate of θ_t, given the data.

Yet, systems like the one from Eq. (ii) exhibit regular, uniform, oscillations, and their phase can be perfectly estimated from e.g. a Fourier transform. In practice, one is often interested in systems showing more complex behaviour, that is, systems under perturbation, or systems of coupled oscillators. Besides, it is often the case that a given oscillator is tracked with different reporters (e.g. the cell-cycle and the 2-colours FUCCI system[311]). A more general model would, therefore, be composed of a state vector $\boldsymbol{\theta} = [\theta_1, \ldots \theta_N]^T$ along with an observations vector $\mathbf{D} = [\mathbf{y}_1, \ldots \mathbf{y}_N]^T$ such that:

$$\begin{cases} d\theta_1 = \omega_{\theta_1} dt + F_1(\boldsymbol{\theta}, \mathrm{t})dt + dW \\ \vdots \\ d\theta_N = \omega_{\theta_N} dt + F_N(\boldsymbol{\theta}, \mathrm{t})dt + dW \end{cases} \quad \text{with} \quad \begin{cases} \mathbf{y}_{1,t} = [s_{1,1}(\theta_1) + \epsilon_{1,1,t}, \ldots, s_{1,M}(\theta_1) + \epsilon_{1,M,t}] \\ \vdots \\ \mathbf{y}_{N,t} = [s_{N,1}(\theta_N) + \epsilon_{N,1,t}, \ldots, s_{N,M}(\theta_N) + \epsilon_{N,M,t}] \end{cases} \qquad (iii)$$

Here, the functions $F_i(\boldsymbol{\theta}, \mathrm{t})$ are considered time-dependent, meaning that the whole dynamical system is not autonomous. In many cases, however, time is not involved, and F_i can be represented on N-dimensional flat toruses, and be parameterized, for instance, using Fourier series. F_i represent the couplings between all the variables in play, along with possible external perturbations. In practice, such couplings rarely link more than two or three variables at a time. As in the uniform 1-dimensional system, the objective here is to obtain the probability of the phases given the data $P(\boldsymbol{\theta}|\mathbf{D})$, or at least a reasonable estimate of $\boldsymbol{\theta}$ given the data.

This is not a simple problem as some, and sometimes all, of the model parameters can be unknown. For instance, intrinsic frequencies can vary with time and be hard to estimate. Similarly, technical noise can be hard to distinguish from phase noise. Finally, in complex molecular systems, the coupling function is usually hard to infer from a mechanistic modelling approach.

Two types of solution have been used in the past. Most of the time, the dynamical phase model is ignored, and phase is inferred directly from the signal with non-parametric approaches. Alternatively, one can take advantage of (part of) the underlying phase model and use a parametric inference approach. In most cases, the latter option has the advantage of yielding a phase probability distribution instead of a single estimate. Besides, one can incorporate prior knowledge about the phase model in the inference process.

2.2.3. Concrete difficulties with phase inference

Phase inference is not a trivial problem for several reasons. First, depending on the shape of the waveform, a given observation y_t can correspond from 2 to n possible phases θ_t, even in the absence of noise. This means that the temporal evolution of the signal must be incorporated in the inference process. At least one full cycle is needed, but this latter condition can also be hard to implement, since identifying a cycle can be difficult in noisy data. Moreover, since phase evolves in a circular space, a large phase deviation can correspond either to a backward or a forward advance, with no way of distinguishing the two using only local information. On longer timescales, signals often contain trends, with amplitude variations, making it difficult to know if a change in signal value is due to to a general trend, or a phase advance. In any case, a way to correct for, or to incorporate, the trend must be implemented in the inference process.

Another recurrent problem with phase inference is waveform identification. The waveform can easily be identified in an isolated system at steady-state, as it should correspond to a noiseless version of a full cycle. However, in the case of a coupled oscillator at steady-state, it is virtually impossible to distinguish the coupling contribution from the natural behaviour of the oscillator. By involving external knowledge about the coupling, one may be able to disentangle what actually belongs to the system natural's dynamics, but this requires a sophisticated parametric approach. In most case, opting for a different experimental design in which the system is first isolated can help to identify an unbiased waveform. The same problem also occurs with the coupling: many interacting oscillators at steady state can be considered as a single oscillatory system, as the respective phases actually follow a limit-cycle in the phase-space. If the system is very noisy, analyzing the phase deviation from the attracting limit-cycle can provide information about the coupling, but this remains a difficult method to implement. Again, in an optimal setting, one would capture the behaviour of isolated oscillators, and only understand *a posteriori* how they interact, precisely by analyzing how the different phases deviate from their natural dynamics when coupled.

2.3. The methods available

2.3.1. Introduction

Despite the development of powerful inference methods in the last fifty years, most studies still use simple linear interpolations between the signal peaks to infer the corresponding phase[312], [313], providing only a zero-order approximation of the phase[303]. Unfortunately, this is imprecise and only provides information on the periods' variation rather than the phase.

Whenever several observation channels are available (e.g. fluorescence microscopy), the inference is sometimes turned into piecewise linear[44]. Although this does bring a non-negligible benefit, the interpolation can be done in several ways depending on the reference points chosen on the signal, and each of these ways provides different results.

Alternatively, if the signal is not too noisy and does not present particular trends, the Hilbert transform is sometimes used. This is, for instance, the case for seismic data[314]. Although it performs very well for sinusoidal data, it is not appropriate for any other waveform. Moreover, it can't handle several channels of observations, limiting its use.

In the general case, that is, a noisy signal with a trend and an arbitrary waveform, with several channels of observations, Bayesian smoothings methods provide the best alternative. The Extended Kalman filter and its extension, the ERTS smoothing, can work very well for simple phase inference, although the non-linearity of the waveform can lead to small biases in the inference[308]. More recently, URTS and particle filtering were developed, allowing to better handle non-linearity of both the phase and observation processes, but at the cost of the optimization of several new parameters used for distribution sampling[308].

Bayesian smoothing methods assume that $P(\theta_t|D)$ is always Gaussian, enabling fast computations since only the means and covariance matrices of the distributions are used to predict and update the distribution at the next timepoint. However, there is one drawback: phase periodicity can't be handled explicitly, since it is assumed that both the state and the observations evolve in R^N. If the signal to noise ratio is high enough, local linearization of the model equation works very well, although the discontinuity between 0 and 2π must still be explicitly considered in the equations. However, the initial estimate is rarely known with certainty, and this can prevent the smoothing of the whole signal. To better handle the domain periodicity, several extensions of these smoothing methods have been developped[315]–[317], using projected Gaussian distributions, but always under the assumption that the observation follows a linear transformation. This is not the case for a waveform, meaning that, to this day, there's no unbiased filter properly handling phase inference from a noisy signal.

Finally, an alternative, simple and exact method for phase inference of any type of signal is the Hidden Markov Model (HMM). Although it also belongs to the family of Bayesian smoothing method, the corresponding phase distribution doesn't suffer from any assumption, meaning that no approximation is made concerning the underlying state process. HMM are easy to implement, and can readily be adapted to work with wrapped normal laws to handle the periodicity of the phase domain[46]. However, contrarily to filtering methods, their tractability can quickly become limiting since the phase inference require the use of several

convolutions on a discretized version of the state space. Moreover, the state space size grows exponentially with the number of hidden variables (phase, and signal trends). This problem can be mitigated by the use of small state space resolution, at the expense of inference precision.

2.3.2. Linear and piecewise linear interpolation

Linear and piecewise linear interpolation are used whenever one is only interested in the average phase deviation over the totality or part of the complete cycle.

Assume the peaks of the signal have been detected at times $[t_1, \dots, t_i, \dots T_n]$. Using linear interpolation, the phase at time $t \in [t_i, t_{i+1})$ is defined as $\theta_t = 2\pi \frac{t-t_i}{t_{i+1}-t_i}$. Of course, any easily identifiable signal point can be used instead of the peaks to obtain the t_i, but using peaks/troughs ensure a proper phase definition since the waveform is unambiguously invertible for these points.

The process with piecewise linear interpolation is slightly more complicated since it requires to define several *turning points* in the observation data. This can be a crossing of the signals coming from the different channels, or possibly their respective peaks, or any other easily identifiable point in each of the signals, taken together or apart. For simplicity, we assume that the signal peaks correspond to one of these turnings points, such that the distribution of time of these turning points is as follows : $[t_1, t_{a_1}, t_{b_1} \dots, t_i, t_{a_i}, t_{b_i} \dots T_n]$, where t_{a_i}, t_{b_i} etc represent the times of the different turning points a_i, b_i etc, with the assumption that $a \equiv a_i (\text{mod } 2\pi), b \equiv b_i (\text{mod } 2\pi), \dots \forall i$ and a, b, ..., represent the (fixed) phases associated with the turning points. The phase at time $t \in [t_i, t_{i+1})$ is then defined as

$$\begin{cases} \theta_t = a \dfrac{t - t_i}{a - t_i} \text{ if } t < t_{a_i} \\ \theta_t = a \dfrac{t - t_i}{a - t_i} + b \dfrac{t - b}{b - a} \text{ if } t_{a_i} \leq t < t_{b_i} \\ \vdots \\ \theta_t = \displaystyle\sum_{x \in a,b,\dots} x \frac{t - t_i}{x - t_i} \text{ else} \end{cases} \qquad (iv)$$

This means that a punctual perturbation of the signal will only be inferred as a variation of the period length, or as a variation of a given time interval between two turning points. In the case of piecewise linear interpolation, the method can be sensitive to the choice of turning points, and, in the absence of control data, there's no way to know if the inference is unbiased. Interpolation methods are therefore imprecise and should be favoured only when punctual phase perturbations are of no interest. Given that the peaks are easily detectable, they are, however, unaffected by signal trends, noise, and fast frequency variation.

2.3.3. Hilbert transform

The Hilbert transform is a quick and efficient method to obtain the phase of a sinusoidal signal. Mathematically, it is defined for a real-valued signal $u(t)$ as:

$$H(u)(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{u(\tau)}{t-\tau}\, d\tau \qquad (v)$$

Although this definition can be interpreted as the convolution of the signal with the function $\frac{1}{\pi t}$, this won't give most readers any insight into what this convolution *actually* does. In practice, it is much easier to understand the Hilbert transform as the imaginary part of the positive frequencies coming from the Fourier transform of the signal. As a reminder, the Fourier Transform of $u(t)$ is defined for frequency f as:

$$\mathcal{F}(u)(f) = \int_{-\infty}^{\infty} u\,(t) e^{-j2\pi f t} dt \qquad (vi)$$

If u is a real-valued signal, then $F(u)$ is necessarily complex-symmetric, meaning that for each positive frequency returned by the transform, there also exists an identical frequency in the negative domain. By removing the redundant negative frequency content, one can create a complex-valued signal whose spectrum is now one-sided, and which preserves the spectral content of the original real-valued signal (by doubling the amplitude of the remaining frequencies). This is the analytic signal, $z(t)$, whose spectral content Z(f) is defined as:

$$Z(f) = \begin{cases} \mathcal{F}(0) & for\ f = 0 \\ 2\mathcal{F}(f) & for\ f > 0 \\ 0 & for\ f < 0 \end{cases} \qquad (vii)$$

Since the spectrum of the analytic signal is asymmetric, the analytic signal is complex-valued in the time domain, and, surprisingly, it is the Hilbert transform that provides the imaginary part of this signal:

$$z(t) = \mathcal{F}^{-1}(Z)(t) = u(t) + iH(u)(t) = A(t)e^{i\theta(t)} \qquad (viii)$$

This means that the instantaneous phase of the signal $u(t)$ can be obtained easily from $\theta(t) = atan2(H(u)(t), u(t))$, while the amplitude is simply $A(t) = |z(t)|$.

It is often said the use of the Hilbert transform should be restricted to a narrowband signal, meaning that the signal should be relatively smooth and untrended. In practice, the Hilbert transform of any real-valued signal is perfectly valid and well-defined at all points, but the returned phase may not correspond to what would intuitively be expected. Indeed, a trended signal can be such that the complex value associated with the trend frequency can completely dominate the spectrum when doing the reversed Fourier transform. The desired phase

oscillates faster than the computed one, and the result is simply not interpretable. The inverse problem is also true: a noise associated with very fast frequencies will also impact the final phase estimate, although in a less important manner. Therefore, one should try to filter all the undesired frequencies from the signal before attempting to make any phase inference with the Hilbert transform. However, in the case of a perturbed signal, the trends may also correspond to real phase deviation and shouldn't be filtered out.

Finally, there exists an extension to the Hilbert transform called the Hilbert-Huang transform[309]. This method first decomposes the signal into different harmonics before applying the Hilbert transform on each of them. It is therefore adapted for data that is non-stationary and non-linear, as the one usually produced by noisy biological oscillators.

2.3.4. Smoothing methods and HMM

Consider a system with N phases to infer from M observations. We denote by $\mathbf{x}_t$ the vector of phase distributions at time t, $\mathbf{q}_t \sim \mathcal{N}(0, \mathbf{Q})$ the corresponding process noise, $\mathbf{y}_t$ the vector of observation and $\mathbf{r}_t$ the measurement noise. The corresponding state-space model is:

$$\begin{cases} \mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}, \boldsymbol{\lambda}) + \mathbf{q}_{k-1} \\ \mathbf{y}_k = \mathbf{h}(\mathbf{x}_k, \boldsymbol{\lambda}) + \mathbf{r}_k \end{cases} \qquad (ix)$$

In this equation, $\mathbf{f}$ is the dynamic model function (e.g. $f = \omega_\theta$ in the case of a single oscillator with linear phase) and $\mathbf{h}$ is the measurement model function, which we previously denoted by waveform. $\boldsymbol{\lambda}$ is the model vector of parameters, tuning features like coupling structure or intensity.

Eq. (ix) is a dynamical system of the form of Eq. (iii). This is precisely the advantage of smoothing methods: they explicitly incorporate a parametric version of the model for which the phase is inferred, and as such, they are much more flexible. They come in various forms, depending on the linearity of the model and measurement functions. Still, in the end, they all return the probability of the phase at any iteration k given the whole data series $P(\mathbf{x}_k, \mathbf{y}_{1:T})$.

The most famous one is also the most basic, known as the Rauch Tung Striebel (RTS) smoother[318]. It is an extension of the Kalman filter, and, as such, only handles linear model and measurement functions. However, by Taylor-expanding the two equations in Eq. (ix), one can approximate reasonably non-linear systems with excellent precision: this is the Extended RTS (ERTS). Another method of linearization, called *unscented*, relies on the use of sigma-points, that is, wisely sampled points to approximate the phase-distribution between two iterations. This gave birth to the Unscented RTS (URTS), which is the most powerful,

but also one the greediest method that we present in this technical review. Finally, Hidden Markov Models (HMM) do not belong to the class of smoothing methods *stricto sensu*, but still constitute the underlying model of all of them. The only difference is that the state space (and also, possibly, the observation space) is discrete. Therefore, while conventional smoothing methods assume continuous, Gaussian distributions, HMM can approximate any system dynamics, whatever the underlying distributions. This is, however, at the expense of tractability: for HMM, computing the state transition between two observations implies convolutions between multidimensional arrays. In contrast, for most smoothing methods, the product of multivariate Gaussians only involves few scalar operations.

The principle of all the smoothing methods and HMM is always the same. First, one assumes an initial probability distribution for the phase, and compute the expected probability of the corresponding observation. By comparing the expected observation with the actual observation in the data, one can correct the phase distribution appropriately. Then, since the process described by Eq. (ix) is Markovian, the phase distribution at the next step can be computed directly from the model function, and same for the corresponding observation. The phase distribution can be corrected, and so on. In practice, one computes iteratively $P(\mathbf{x}_1, \mathbf{y}_1)$,$P(\mathbf{x}_2, \mathbf{y}_{1:2}), \ldots, P(\mathbf{x}_T, \mathbf{y}_{1:T})$. This is known as the *filtering* step, or forward pass for HMM. From reference [319], the algorithm principle is as follow:

Initialization

1. Initialize the state of the filter

2. Initialize our belief in the state

Prediction

1. Use system behaviour to predict state at the next time step

2. Adjust belief to account for the uncertainty in prediction

Update

1. Get a measurement and associated belief about its accuracy

2. Compute residual between estimated state and measurement

3. Compute scaling factor based on whether the measurement or prediction is more accurate

4. set state between the prediction and measurement based on scaling factor

5. update belief of the state based on how certain we are in the measurement

The *smoothing* step, or backward pass for HMM, that comes afterwards is more of a correction. Indeed, the filtering step only provides $P(\mathbf{x}_k, \mathbf{y}_{1:k})$, that is, not all data is incorporated in the estimate. By processing the data in reverse order, the smoothing step yields $P(\mathbf{x}_k, \mathbf{y}_{k+1:T})$, from which one can compute $P(\mathbf{x}_k, \mathbf{y}_{1:T})$.

The advantage of smoothing methods is that they work with any type of waveform, they can explicitly account for coupling of as many oscillators as needed, and also account for any trend in the data. The disadvantages are that the structure of the phase and observation models must be at least partly known to be optimized, and that the whole process is much greedier than non-parametric methods such as the Hilbert transform. Still, the worst tractability belongs to HMM, for which the size of the state space grows exponentially with the resolution chosen, as well with the number of variables.

2.4. Method Performances

To compare the performance of the methods presented above, we simulate traces using Eq. (iii), considering only two phases, θ and ϕ, along with various choices of waveforms (Figure A.1A) and coupling functions (Figure A.1B-C). From here, one can generate the successive phase states, along with the corresponding signal observations (Figure A.1D-E).

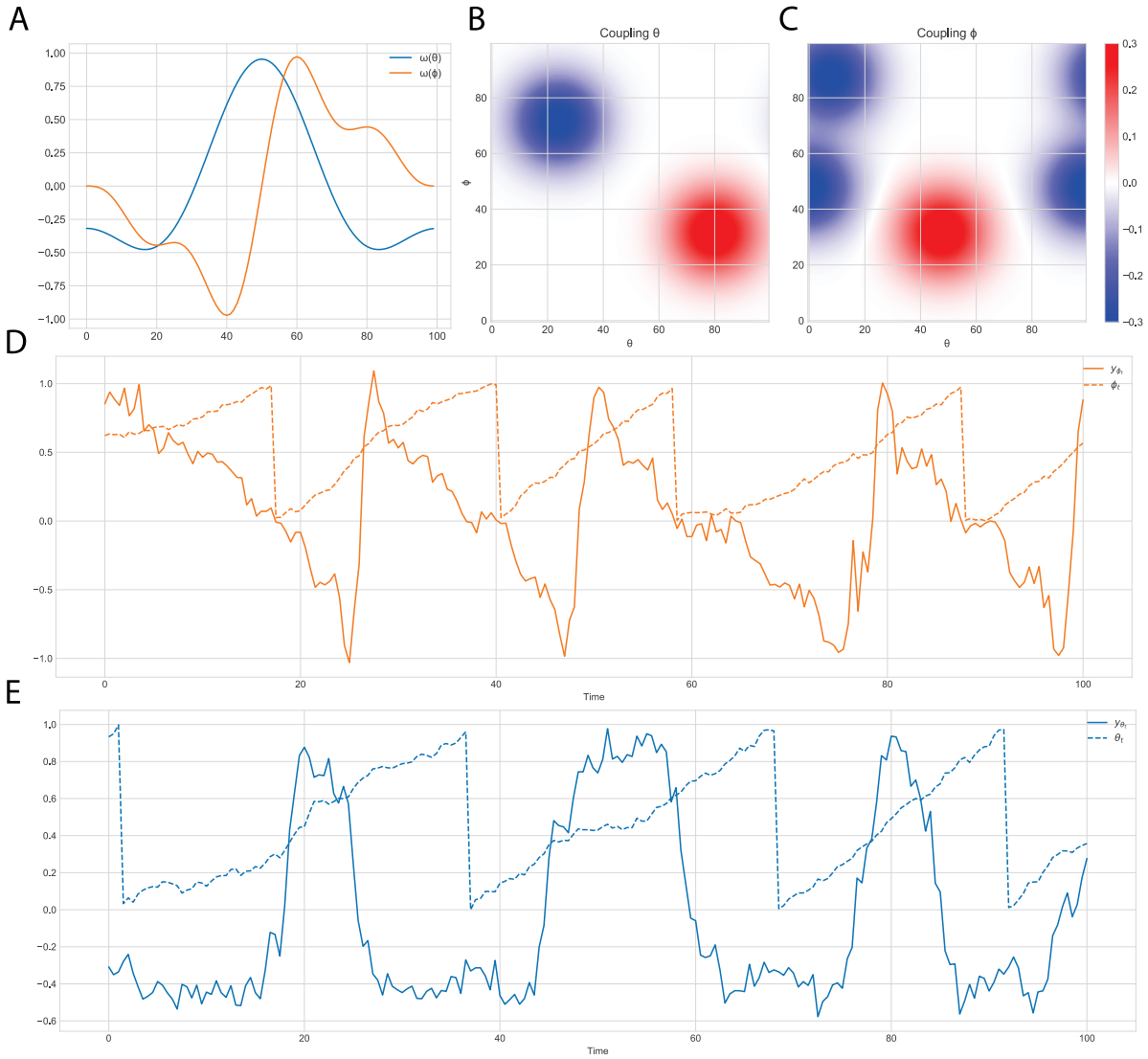


Figure A.1: Data simulation. (**A**) The two waveforms used to generate the observations presented in D-E. (**B-C**) The corresponding coupling function, used in the dynamic model function, to modulate the transition between the phase states. (**D-E**) Temporal evolution of the phase (dashed line) along with the corresponding observations (full line). Observation noise is Gaussian with mean 0.15, phase noise is Gaussian with mean 0.1.

One can then run the inference and, by computing the two-dimensional inferred phase discrete derivatives, get an approximation of the coupling functions (Figure A.2). None of the methods perform well enough to reconstitute an accurate description of the real coupling, but all of the methods presented yield a qualitatively correct approximation. Using a regularization process could here help with the noisy area of the phase-space. Except if vast amounts of data are available, coupling inference is not really applicable to linear interpolation methods as they necessarily predict a flat coupling for a given cycle. Piecewise linear interpolations can perform slightly better (cf. the work done by *Feillet et al.* [44]) but are still much below other methods in terms of accuracy.

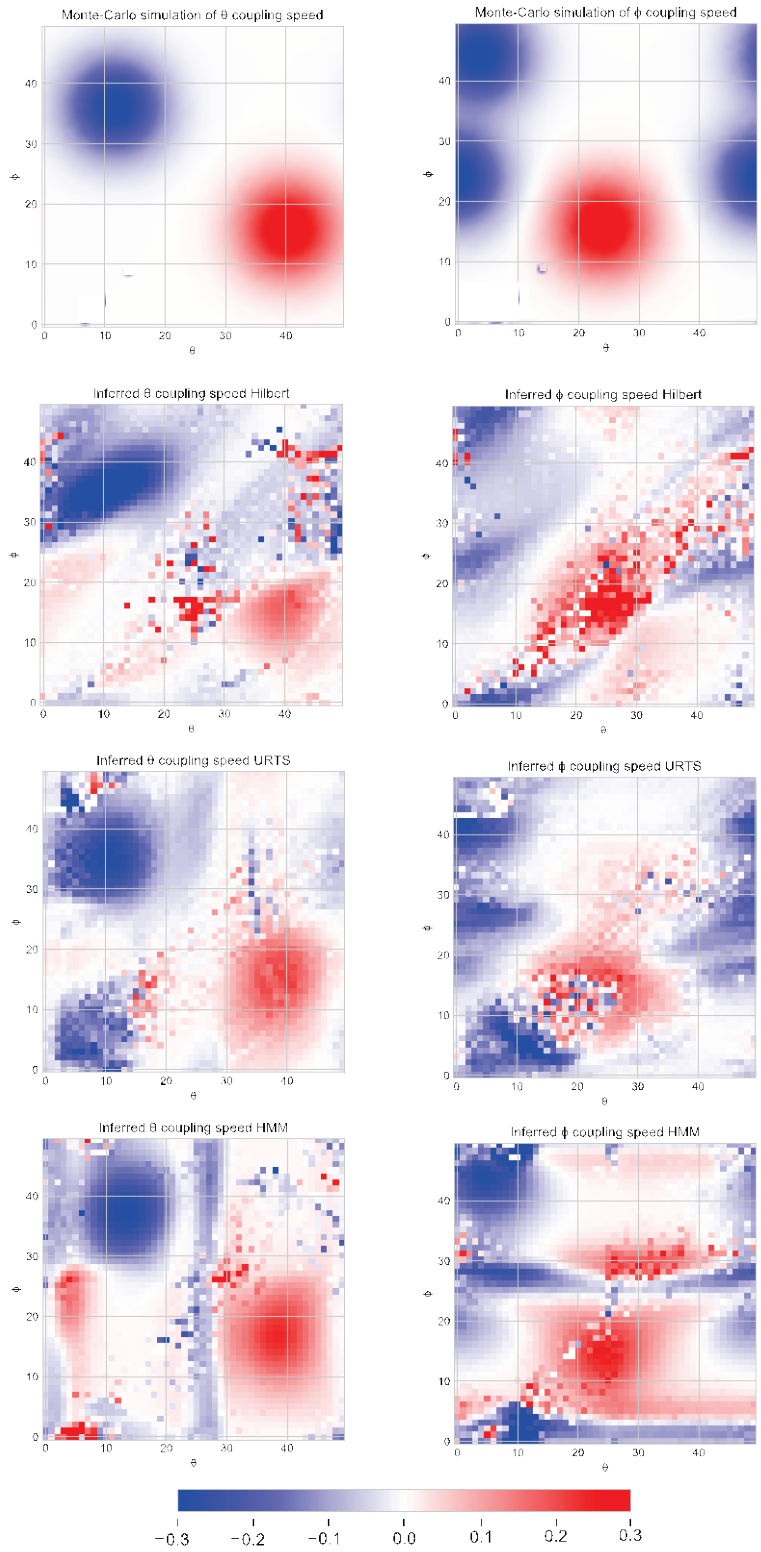


Figure A.2: Comparison of the inferred coupling depending on the phase inference method for both the $\boldsymbol{\theta}$ (left) and $\boldsymbol{\varphi}$ (right) coupling functions. The coupling is computed as the average discrete derivatives of $\boldsymbol{\theta}$ (**left**) and $\boldsymbol{\varphi}$ (**right**), with respect to time, for each possible couple of phases. In order, from top to bottom, the represented couplings are as follow: True coupling, coupling inferred from Hilbert transform, coupling inferred from URTS, coupling inferred from HMM.

It's also interesting to take a look at how the different methods perform at the individual signal level (Figure A.3). Even on a noiseless signal, substantial discrepancies can occur between the different methods.

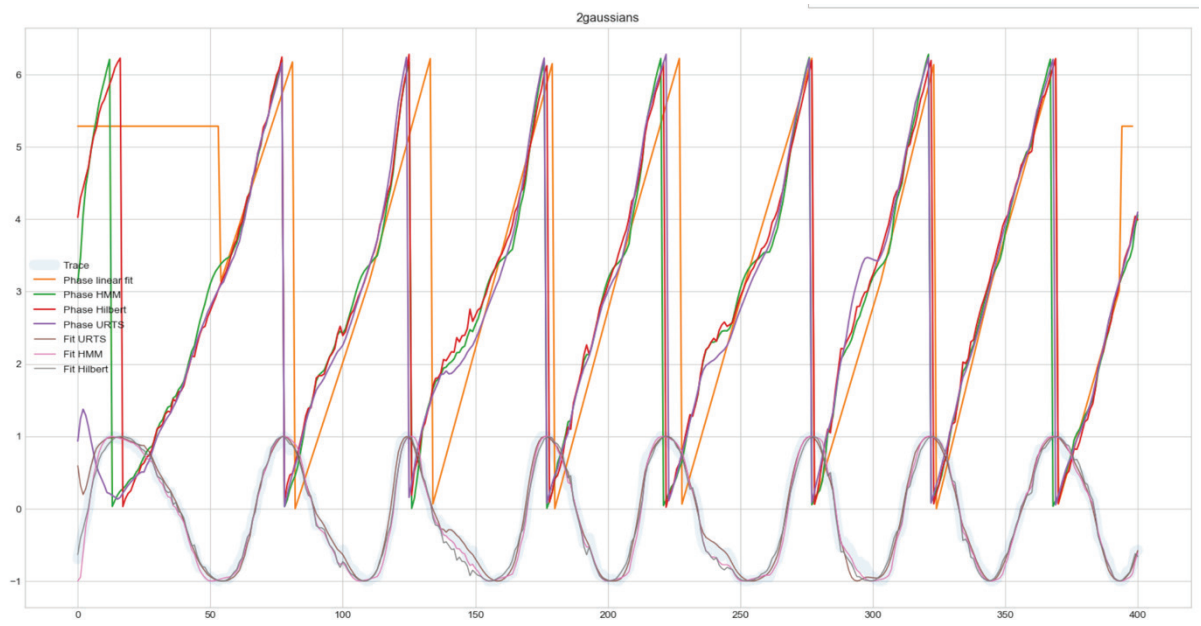


Figure A.3: Comparison of the inferred phases with the different methods presented above. The simulated data ("trace") is based on a sinusoidal signal under external forcing.

2.5. Code availability

This whole material used to generate this review, including the implementation of each method, is available online as a collection of Julia Jupyter notebooks at https://github.com/ColasDroin/PhaseInferenceReview.

3. Perspectives

This study is only a draft, and much remains to be done. Although the implementation of the different methods was undoubtedly the hardest part (8 classes implemented, several thousands line of code), finding a good measure to compare the different methods in all possible conditions is not an easy task. This is because many variables are present in the problem: the type of waveform, of coupling, of intrinsic and extrinsic noise, the number of oscillators, the dynamical changes of frequency, etc. Therefore, it is a complex problem to summarize all these conditions into a single table.

Obviously, one should also test the different methods on real datasets. FUCCI signals would make excellent candidates as they use several channels (discarding the Hilbert-transform) and are widespread in cell-cycle studies. Ideally, one would also study data coming from neuroscience studies, as they involve complex couplings between different areas of the brain.

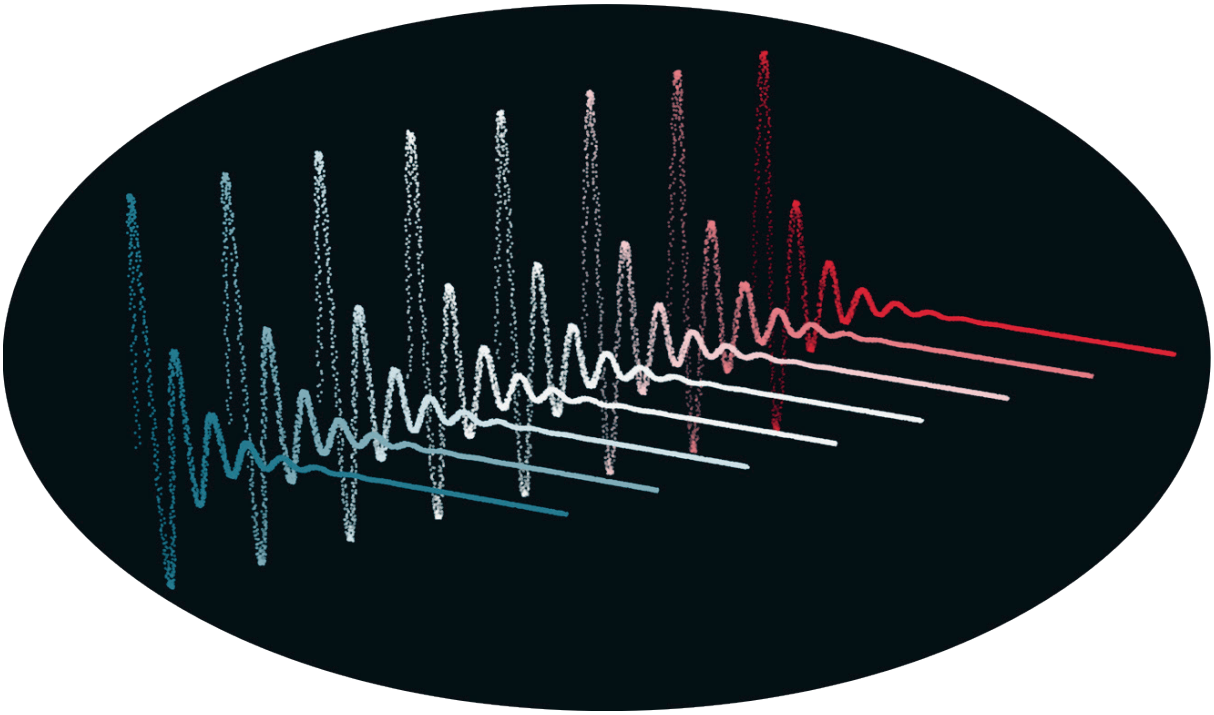
Finally, it would be interesting would like to explore one of the most powerful black-box methods out there: the Hilbert-Huang transform, which, *a priori*, could yield results just as good as smoothing methods, for a fraction of the computational power.

Annexe B: Widgets and animations

The chapter that follows presents the design and implementation of several widgets and animations used to explain phenomena related to phase inference and phase dynamics.

Contributions

Under the supervision of F. Naef, I designed all the widgets and animations, and implemented the corresponding code.



Artwork Figure 7: Artistic representation of noisy decaying oscillations, slightly dephased on the y-axis (depth). This plot is based on a modified simulation of fluorescence microscopy experiments in which the circadian clocks of single cells become progressively desynchronized.

1. Introduction

All the widgets and animations that are presented in the following subsections have been developed in the context of my PhD, to explain or clarify concepts for which sole equations can be hard to grasp. Except for the Fourier decomposition widget, which uses D3.js, they are all implemented in Python using the Matplotlib library. The code is open-source, freely available on Github at https://github.com/ColasDroin/Python-animations and https://github.com/ColasDroin/D3-Fourier.

2. Phase-space animation

This set of animation was designed to provide an intuitive explanation of the behaviour of coupled oscillators systems. The very first animation that I created (Figure B.1) was thought to support our coupling paper[46]: it represents the time evolution of stochastic trajectories in the phase space of the cell-cycle and the circadian clock, with the corresponding inferred coupling function in the background. Trajectories are inferred from real NIH3T3 fluorescent traces, with randomly distributed colours. Due to the dominance of the 1:1 mode-locking in the system, trajectories tend to have a slope of about one, gathering between the accelerating (red) and decelerating (blue) regions of the phase-space, although the noise can sometimes kick them out (cf. Introduction, Section 3.4.1.2 of this thesis).

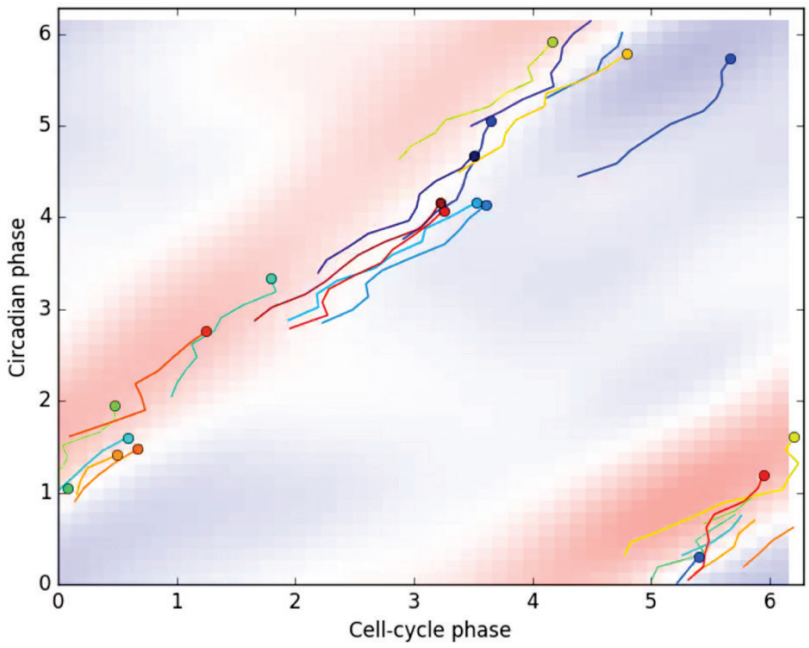


Figure B.1: Snapshot of an animation used to represent the time-evolution of 20 NIH3T3 inferred trajectories in the phase-space of the cell-cycle and the circadian clock. The background represents the coupling function, with red (blue) areas accelerating (decelerating) the circadian clock. The instantaneous value of the cells is indicated by a coloured disc, while the previous values are queuing behind (10 timepoints, corresponding to 5h, are kept). Coloured are assigned randomly.

The second animation (Figure B.2) has been developed about two years later, while the coupling study was being published. It was thought to explain the behaviour of noisy phase-locked systems. As previously, the temporal trajectory of two oscillators with phase ϕ_1 and ϕ_2 is represented in the phase-space, with the coupling function in the background (red and blue Gaussians, representing accelerating and decelerating areas for ϕ_2, respectively). However, contrarily to Figure B.1, there's only one trajectory, which is simulated and imprints a track on the phase-space, enabling to highlight the existence of a mode-locking (2:1 here, but other modes are possible in different versions of the animation). To make the idea of boundary periodicity more intuitive, the coupling function is also represented on a three-dimensional rotating torus (Figure B.3).

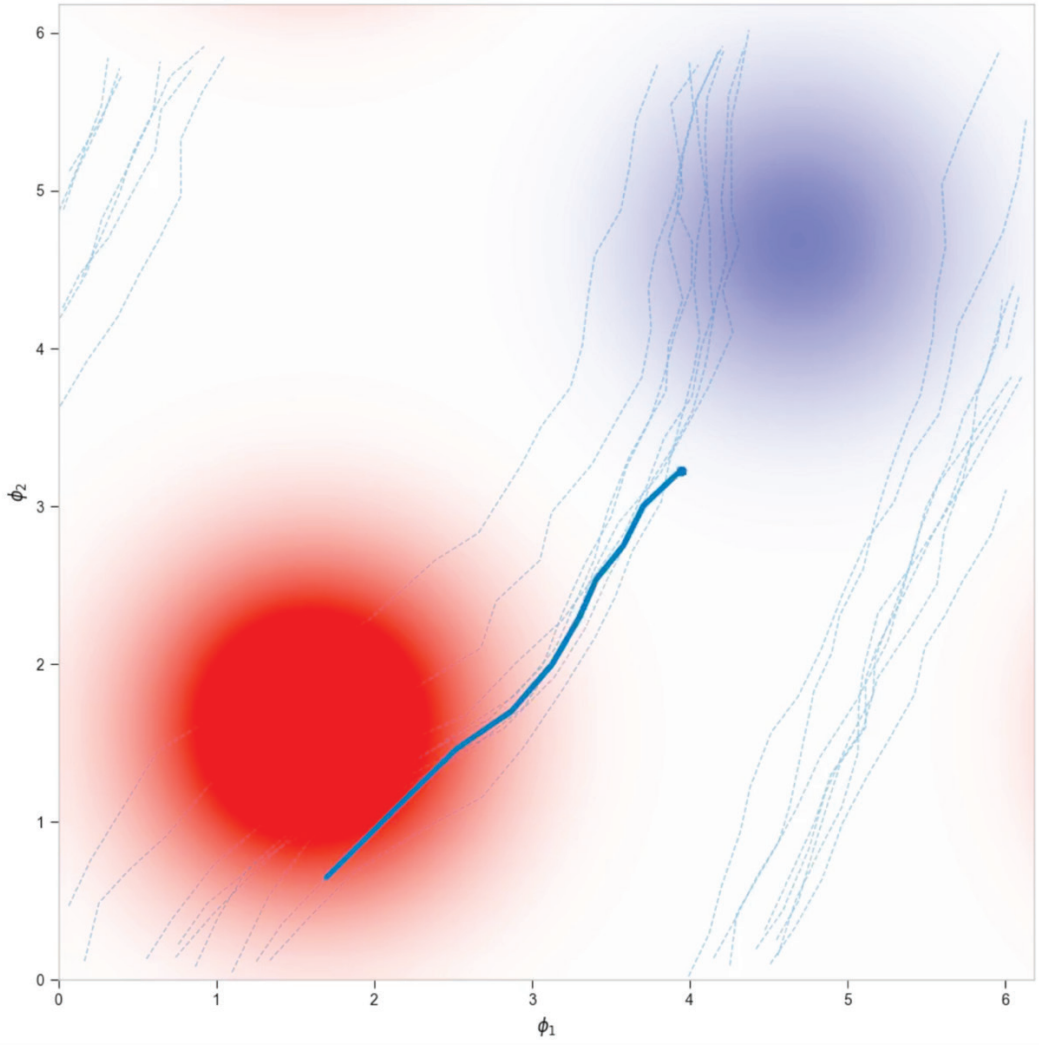


Figure B.2: Snapshot of an animation used to represent the time-evolution of 1 simulated trajectory in the phase-space of two oscillators with phases ϕ_1 and ϕ_2. The background represents the coupling function, with red (blue) areas accelerating (decelerating) ϕ_2. The instantaneous value of the cells is indicated by a coloured disc, while the queue indicates the previous timepoints. This animation enables to highlight the presence of a 2:1 mode-locking the system: ϕ_2 makes two full cycles when ϕ_1 does one, even though the system is stochastic. Simulations are computed using a Monte-Carlo approach.

Figure B.3: Snapshot of an animation representing the coupling function from Figure B.2, imprinted on a rotating torus. Red (blue) areas accelerate (decelerate) the phase of oscillator 1.

The widget presented Figure B.4 is more interactive and pedagogic than the previous ones. It represents the progression of the phase of two oscillators on their respective waveform (left, just a sine function here). The progression in the phase-space is represented on the bottom-right panel, with the coupling as background (same colour-code as before). Finally, the dephasing (top right) for the last 20 timepoints is also indicated, revealing how the two oscillators tend to stabilize around a given phase difference for a given coupling. Although the system quickly reaches a steady-state in which not much happens, the user can change the strength of the coupling in real time (bottom bar, K), and observe the system's transient behaviour as time evolves (bottom bar, t).

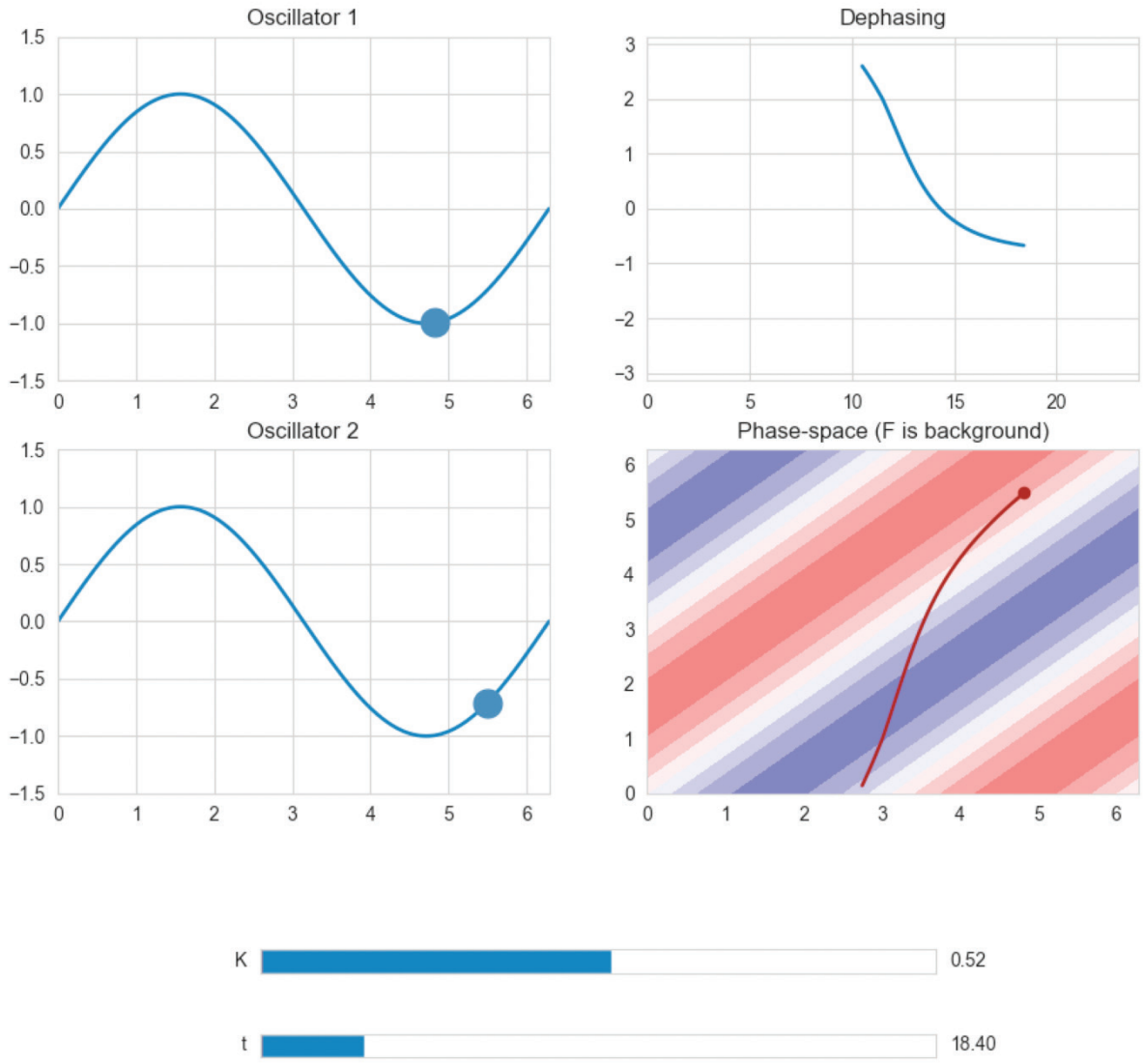


Figure B.4: Snapshot of an interactive Python widget, in which two oscillators interact. The phase progression along the waveform of the two oscillators is represented on the left (top and bottom). Their dephasing is represented on the top-right corner. Their trajectory in the phase-space is represented on the bottom-right corner, with the coupling function as background (red (blue) areas accelerating (decelerating) oscillator 2. The two interactive bars in the bottom represent the coupling strength (K) and the time (t).

Finally, to provide an intuitive idea of what the model used in the coupling study represents, I've developed an animated version of the sampled signal observations (Figure B.5). The model is as follow:

$$\begin{cases} \mathrm{d}\theta_t = \dfrac{2\pi}{T_\theta}\mathrm{d}t + \sigma_\theta\,\mathrm{d}W_{\theta,t} \\ \mathrm{d}A_t = -\gamma_A(A_t - \mu_A)dt + \sigma_A\,dW_{A,t} \\ \mathrm{d}B_t = -\gamma_B(B_t - \mu_B)\mathrm{d}t + \sigma_B\,\mathrm{d}W_{B,t} \\ S_t = \exp(A_t)w(\theta_t) + B_t + \xi. \end{cases} \qquad (i)$$

The corresponding animation starts from the simplest possible version of Eq. (i), that is, no amplitude fluctuations, no background, no noise, and, after several full-cycles, progressively incorporates the zeroed-out terms until the final model is reached again.

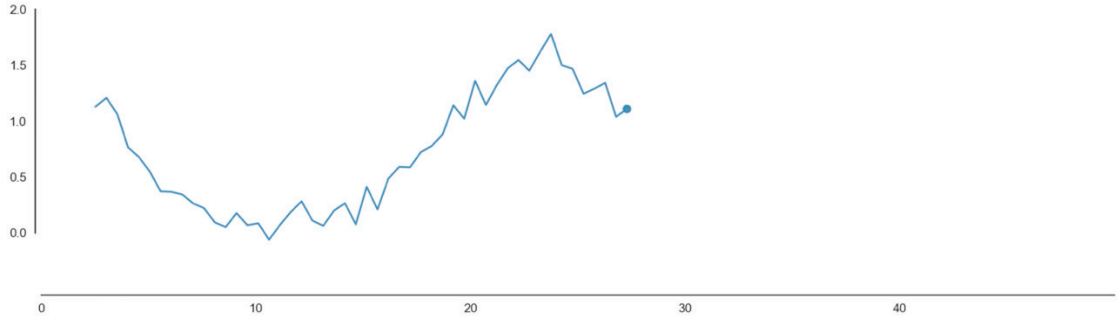


Figure B.5: Snapshot of an animation representing a simulation of Eq. (i). The emission of the system at time t is represented by the leading blue point, while the trail behind represents the previous observations, during the 50 last timepoints). The animation is made such that the system starts from a simplified version of Eq. (i), with no noise, no amplitude, and no background, and gets progressively complexified until the final version of the model is reached.

3. Oscillator time trajectories

This animation (Figure B.6) has been developed to clarify the concept of limit-cycle, taking the example of the circadian clock. In practice, bulk mRNA expression data from the core-clock genes is extracted from the *Atger et al.* [227] study and fitted with simple harmonic functions. The fact that the curves only represent the last 12h enables to highlight the 24h periodicity of the clock.

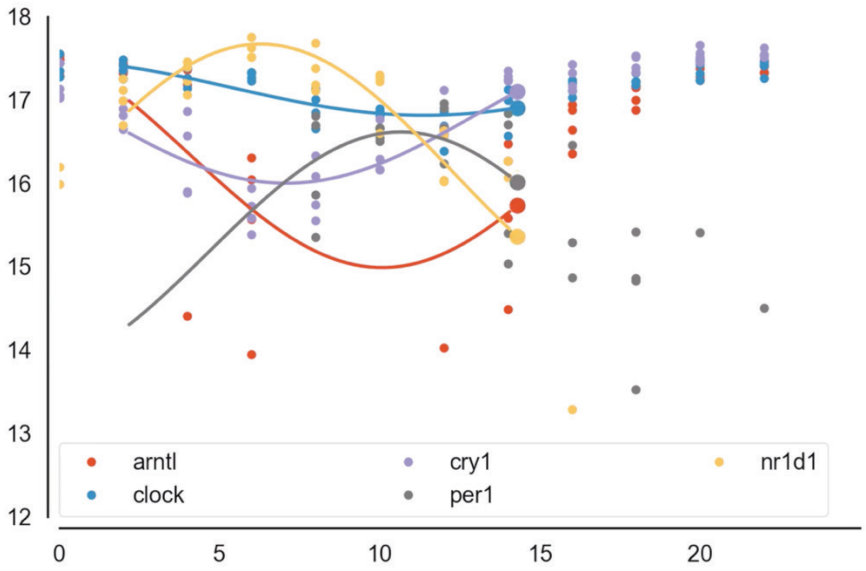


Figure B.6: Snapshot of an animation representing the temporal evolution of mRNA core-clock genes expression. Raw data is scattered (static), while the corresponding fits, coming from simple harmonic regressions, are plotted as time-evolving curves, with a trail of about 12h.

4. Enrichment around the clock

This animation (Figure B.7) has been developed to illustrate the functional analysis of the mouse liver transcriptome around the clock. Functions which are highly enriched (e.g. protein folding here) are selected, and the enrichment is computed around the clock with several hypergeometric tests (cf. Supplementary Information from study [320]). If the enrichment is time-specific, an elliptic shape usually appears (e.g. in purple here). In this animation, a total of 8 functions are iteratively represented, each being highly specific to a given zeitgeber.

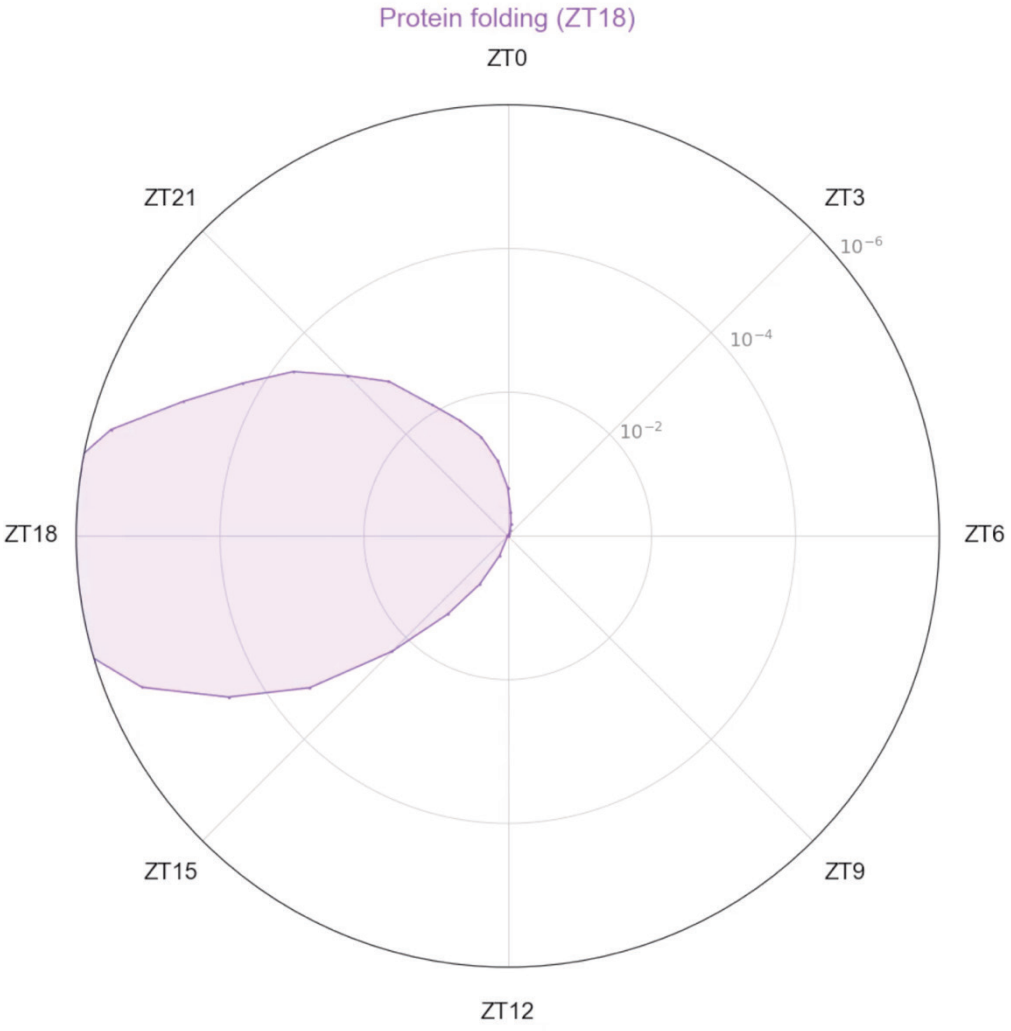


Figure B.7: Snapshot of an animation representing the enrichment around the clock of the mouse transcriptome in genes belonging to preliminarily selected biological functions (here, in purple, protein folding). For visual purposes, transitions between the different functions are interpolated, ellipses are smoothed, and functions are selected to cover the day uniformly. The p-value corresponding to the enrichment is indicated on the radial axis.

5. D3 widget to compute Fourier transform of a signal

This widget (Figure B.8) was initially developed to turn the study presented in Annexe A in what is called an observable study[26]. The principle is the following: the user can draw any curve-like shape on the top window, with the restriction that the result must be a function (there's only one y corresponding to a given x). The drawn curve is then decomposed into a set of harmonics, represented below. In parallel, it is represented as a smooth oscillatory signal in the window below, taking only the first harmonics[27]. The whole widget was implemented in Javascript, with the help of the D3.js library.

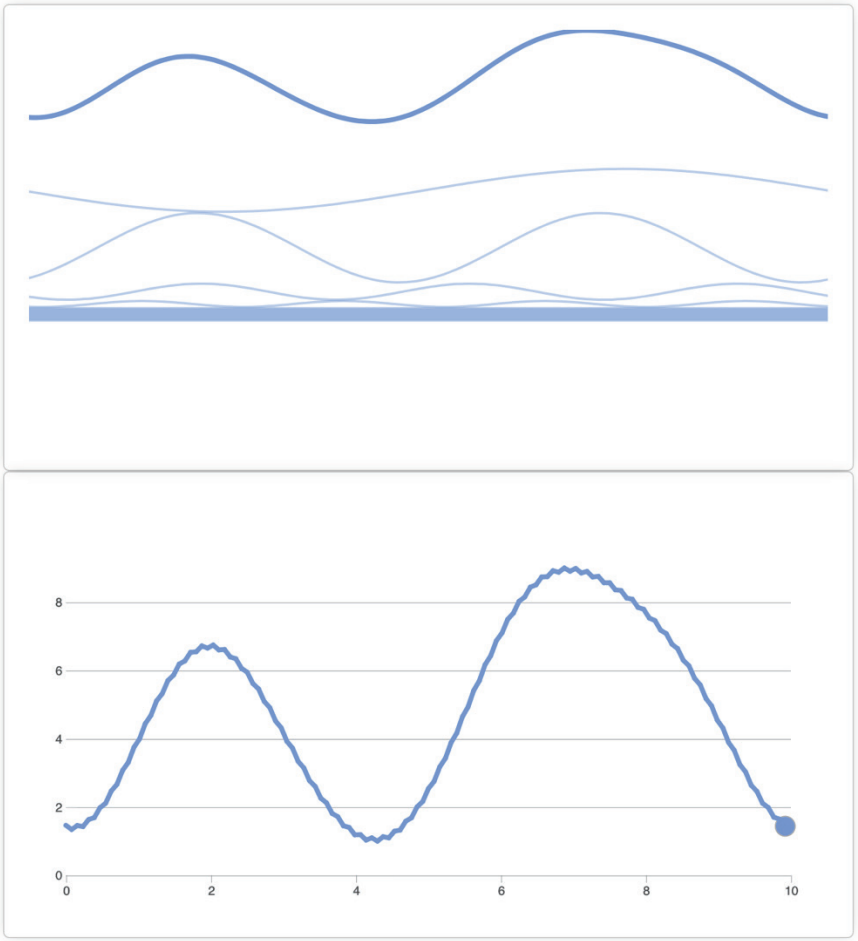


Figure B.8: Snapshot of an interactive widget used to clarify the concept of Fourier transform. In the top window, the user can draw any function with his mouse, which is then decomposed into a sum of Fourier harmonics, represented below. The drawn function is then smoothed (zeroing-out the fastly-cycling harmonics), and represented as a periodic signal in the window below.

[26] See https://observablehq.com/ and https://distill.pub/.

[27] This was implemented with the objective of showing phase inference for the corresponding signal in parallel, but it was left as it as, as the study was dropped.

References

[1] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization*. Cambridge University Press (CUP), 2001.

[2] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*, Second edition. Boulder, CO: Westview Press, a member of the Perseus Books Group, 2015.

[3] D. B. Forger, *Biological clocks, rhythms, and oscillations: the theory of biological timekeeping*. Cambridge, Massachusetts: The MIT Press, 2017.

[4] U. Alon, *An introduction to systems biology: design principles of biological circuits*, Second edition. Boca Raton, Fla: CRC Press, 2019.

[5] V. Kumar, *Biological timekeeping: clocks, rhythms and behaviour*. New York, NY: Springer Berlin Heidelberg, 2017.

[6] G. Dupont, L. Combettes, G. S. Bird, and J. W. Putney, 'Calcium Oscillations', *Cold Spring Harb. Perspect. Biol.*, vol. 3, no. 3, pp. a004226–a004226, Mar. 2011, doi: 10.1101/cshperspect.a004226.

[7] J. E. Ferrell, T. Y.-C. Tsai, and Q. Yang, 'Modeling the Cell Cycle: Why Do Certain Circuits Oscillate?', *Cell*, vol. 144, no. 6, pp. 874–885, Mar. 2011, doi: 10.1016/j.cell.2011.03.006.

[8] G. Lorenzi-Filho, H. R. Dajani, R. S. T. Leung, J. S. Floras, and T. D. Bradley, 'Entrainment of Blood Pressure and Heart Rate Oscillations by Periodic Breathing', *Am. J. Respir. Crit. Care Med.*, vol. 159, no. 4, pp. 1147–1154, Apr. 1999, doi: 10.1164/ajrccm.159.4.9806081.

[9] B. Hannon and M. Ruth, 'Modeling Dynamic Biological Systems', in *Modeling Dynamic Biological Systems*, Cham: Springer International Publishing, 2014, pp. 3–28.

[10] P. E. Parham and E. Michael, 'Outbreak properties of epidemic models: The roles of temporal forcing and stochasticity on pathogen invasion dynamics', *J. Theor. Biol.*, vol. 271, no. 1, pp. 1–9, Feb. 2011, doi: 10.1016/j.jtbi.2010.11.015.

[11] S. E. Cohen and S. S. Golden, 'Circadian Rhythms in Cyanobacteria', *Microbiol. Mol. Biol. Rev.*, vol. 79, no. 4, pp. 373–385, Dec. 2015, doi: 10.1128/MMBR.00036-15.

[12] G. Tiana, S. Krishna, S. Pigolotti, M. H. Jensen, and K. Sneppen, 'Oscillations and temporal signalling in cells', *Phys. Biol.*, vol. 4, no. 2, pp. R1–R17, May 2007, doi: 10.1088/1478-3975/4/2/R01.

[13] S. M. Boker, E. Leibenluft, P. R. Deboeck, G. Virk, and T. T. Postolache, 'Mood Oscillations and Coupling Between Mood and Weather in Patients with Rapid Cycling Bipolar Disorder', *Int. J. Child Health Hum. Dev. IJCHD*, vol. 1, no. 2, pp. 181–203, 2008.

[14] A. A. Andronov, A. A. Vitt, S. Ė. Khaĭkin, and W. Fishwick, *Theory of oscillators*. New York: Dover, 1987.

[15] J. R. Westra, C. J. M. Verhoeven, and A. H. M. van Roermund, *Oscillators and Oscillator Systems: Classification, Analysis and Synthesis*. 1999.

[16] *The Geometry of Biological Time*. New York, NY: Springer New York, 2001.

[17] A. L. Hodgkin and A. F. Huxley, 'A quantitative description of membrane current and its application to conduction and excitation in nerve', *J. Physiol.*, vol. 117, no. 4, pp. 500–544, Aug. 1952, doi: 10.1113/jphysiol.1952.sp004764.

[18] I. R. Titze, 'The physics of small-amplitude oscillation of the vocal folds', *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1536–1552, Apr. 1988, doi: 10.1121/1.395910.

[19] N. Mrosovsky, 'Hibernation and body weight in dormice: a new type of endogenous cycle', *Science*, vol. 196, no. 4292, pp. 902–903, May 1977, doi: 10.1126/science.860123.

[20] J. Keener and J. Sneyd, Eds., *Mathematical Physiology*, vol. 8/1. New York, NY: Springer New York, 2009.

[21] J. J. Tyson, 'Biochemical Oscillations', in *Computational Cell Biology*, vol. 20, C. P. Fall, E. S. Marland, J. M. Wagner, and J. J. Tyson, Eds. New York, NY: Springer New York, 2004, pp. 230–260.

[22] K. Uriu, 'Genetic oscillators in development', *Dev. Growth Differ.*, vol. 58, no. 1, pp. 16–30, Jan. 2016, doi: 10.1111/dgd.12262.

[23] D. E. Nelson *et al.*, 'Oscillations in NF-kappaB signaling control the dynamics of gene expression', *Science*, vol. 306, no. 5696, pp. 704–708, Oct. 2004, doi: 10.1126/science.1099962.

[24] Z. Li and Q. Yang, 'Systems and synthetic biology approaches in understanding biological oscillators', *Quant. Biol.*, vol. 6, no. 1, pp. 1–14, Mar. 2018, doi: 10.1007/s40484-017-0120-7.

[25] B. Novák and J. J. Tyson, 'Design principles of biochemical oscillators', *Nat. Rev. Mol. Cell Biol.*, vol. 9, no. 12, pp. 981–991, Dec. 2008, doi: 10.1038/nrm2530.

[26] A. T. Winfree, 'The prehistory of the Belousov-Zhabotinsky oscillator', *J. Chem. Educ.*, vol. 61, no. 8, p. 661, Aug. 1984, doi: 10.1021/ed061p661.

[27] M. H. Vitaterna, J. S. Takahashi, and F. W. Turek, 'Overview of circadian rhythms', *Alcohol Res. Health J. Natl. Inst. Alcohol Abuse Alcohol.*, vol. 25, no. 2, pp. 85–93, 2001.

[28] P. K. Jackson, 'The Hunt for Cyclin', *Cell*, vol. 134, no. 2, pp. 199–202, Jul. 2008, doi: 10.1016/j.cell.2008.07.011.

[29] C. Bodenstein, I. Heiland, and S. Schuster, 'Temperature compensation and entrainment in circadian rhythms', *Phys. Biol.*, vol. 9, no. 3, p. 036011, Jun. 2012, doi: 10.1088/1478-3975/9/3/036011.

[30] M. von Schantz, 'Natural Variation in Human Clocks', in *Advances in Genetics*, vol. 99, Elsevier, 2017, pp. 73–96.

[31] C. R. McClung, 'Plant Circadian Rhythms', *Plant Cell*, vol. 18, no. 4, pp. 792–803, Apr. 2006, doi: 10.1105/tpc.106.040980.

[32] A. Pikovsky and Y. Maistrenko, Eds., *Synchronization: Theory and Application*. Dordrecht: Springer Netherlands, 2003.

[33] L. Glass and M. C. Mackey, *From clocks to chaos: the rhythms of life*. Princeton, N.J: Princeton University Press, 1988.

[34] A. Ramkisoensing and J. H. Meijer, 'Synchronization of Biological Clock Neurons by Light and Peripheral Feedback Systems Promotes Circadian Rhythms and Health', *Front. Neurol.*, vol. 6, Jun. 2015, doi: 10.3389/fneur.2015.00128.

[35] J. J. Hopfield and A. V. Herz, 'Rapid local synchronization of action potentials: toward computation with coupled integrate-and-fire neurons.', *Proc. Natl. Acad. Sci.*, vol. 92, no. 15, pp. 6655–6662, Jul. 1995, doi: 10.1073/pnas.92.15.6655.

[36] L. Weller, A. Weller, H. Koresh-Kamin, and R. Ben-Shoshan, 'Menstrual synchrony in a sample of working women', *Psychoneuroendocrinology*, vol. 24, no. 4, pp. 449–459, May 1999, doi: 10.1016/S0306-4530(98)00092-4.

[37] S. H. Strogatz, 'Spontaneous synchronization in nature', in *Proceedings of International Frequency Control Symposium*, Orlando, FL, USA, 1997, pp. 2–4, doi: 10.1109/FREQ.1997.638513.

[38] Y.-F. Xiao, 'Cardiac arrhythmia and heart failure: From bench to bedside', *J. Geriatr. Cardiol. JGC*, vol. 8, no. 3, pp. 131–132, Sep. 2011, doi: 10.3724/SP.J.1263.2011.00131.

[39] V. Blakeman, J. L. Williams, Q.-J. Meng, and C. H. Streuli, 'Circadian clocks and breast cancer', *Breast Cancer Res.*, vol. 18, no. 1, p. 89, Dec. 2016, doi: 10.1186/s13058-016-0743-z.

[40] L. Nashef, S. Garner, J. W. A. S. Sander, D. R. Fish, and S. D. Shorvon, 'Circumstances of death in sudden death in epilepsy: interviews of bereaved relatives', *J. Neurol. Neurosurg. Psychiatry*, vol. 64, no. 3, pp. 349–352, Mar. 1998, doi: 10.1136/jnnp.64.3.349.

[41] D. J. Watts and S. H. Strogatz, 'Collective dynamics of "small-world" networks', *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998, doi: 10.1038/30918.

[42] A. E. Pereda, 'Electrical synapses and their functional interactions with chemical synapses', *Nat. Rev. Neurosci.*, vol. 15, no. 4, pp. 250–263, Apr. 2014, doi: 10.1038/nrn3708.

[43] S. A. Brown, 'Circadian clock-mediated control of stem cell division and differentiation: beyond night and day', *Development*, vol. 141, no. 16, pp. 3105–3111, Aug. 2014, doi: 10.1242/dev.104851.

[44] C. Feillet *et al.*, 'Phase locking and multiple oscillating attractors for the coupled mammalian clock and cell cycle.', *Proc Natl Acad Sci U A*, vol. 111, pp. 9828–33, Jul. 2014.

[45] J. Bieler, R. Cannavo, K. Gustafson, C. Gobet, D. Gatfield, and F. Naef, 'Robust synchronization of coupled circadian and cell cycle oscillators in single mammalian cells', *Mol. Syst. Biol.*, vol. 10, no. 7, p. 739, Jul. 2014, doi: 10.15252/msb.20145218.

[46] C. Droin, E. R. Paquet, and F. Naef, 'Low-dimensional dynamics of two coupled biological oscillators', *Nat. Phys.*, vol. 15, no. 10, pp. 1086–1094, Oct. 2019, doi: 10.1038/s41567-019-0598-1.

[47] P. A. Tass, *Phase resetting in medicine and biology: stochastic modelling and data analysis*. Berlin ; New York: Springer Verlag, 1999.

[48] D. A. Rand, B. V. Shulgin, J. D. Salazar, and A. J. Millar, 'Uncovering the design principles of circadian clocks: Mathematical analysis of flexibility and evolutionary goals', *J. Theor. Biol.*, vol. 238, no. 3, pp. 616–635, Feb. 2006, doi: 10.1016/j.jtbi.2005.06.026.

[49] *Principles and Practice of Sleep Medicine*. Elsevier, 2017.

[50] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, vol. 42. New York, NY: Springer New York, 1983.

[51] W. O. Friesen and G. D. Block, 'What is a biological oscillator?', *Am. J. Physiol.*, vol. 246, no. 6 Pt 2, pp. R847-853, Jun. 1984, doi: 10.1152/ajpregu.1984.246.6.R847.

[52] B. Cahlon and D. Schmidt, 'Generic Oscillations for Delay Differential Equations', *J. Math. Anal. Appl.*, vol. 223, no. 1, pp. 288–301, Jul. 1998, doi: 10.1006/jmaa.1998.5979.

[53] K. Burrage, J. Hancock, A. Leier, and D. V. Nicolau, 'Modelling and simulation techniques for membrane biology', *Brief. Bioinform.*, vol. 8, no. 4, pp. 234–244, Mar. 2007, doi: 10.1093/bib/bbm033.

[54] S. E. Jørgensen and B. D. Fath, *Encyclopedia of ecology*. Oxford: Elsevier, 2008.

[55] G. Lahav, 'Oscillations by the p53-Mdm2 feedback loop', *Adv. Exp. Med. Biol.*, vol. 641, pp. 28–38, 2008, doi: 10.1007/978-0-387-09794-7_2.

[56] Q. Li and X. Lang, 'Internal noise-sustained circadian rhythms in a Drosophila model', *Biophys. J.*, vol. 94, no. 6, pp. 1983–1994, Mar. 2008, doi: 10.1529/biophysj.107.109611.

[57] N. Geva-Zatorsky, E. Dekel, E. Batchelor, G. Lahav, and U. Alon, 'Fourier analysis and systems identification of the p53 feedback loop', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 30, pp. 13550–13555, Jul. 2010, doi: 10.1073/pnas.1001107107.

[58] D. Gonze and P. Ruoff, 'The Goodwin Oscillator and its Legacy', *Acta Biotheor.*, Mar. 2020, doi: 10.1007/s10441-020-09379-8.

[59] D. Gonze and W. Abou-Jaoudé, 'The Goodwin Model: Behind the Hill Function', *PLoS ONE*, vol. 8, no. 8, p. e69573, Aug. 2013, doi: 10.1371/journal.pone.0069573.

[60] M. B. Elowitz and S. Leibler, 'A synthetic oscillatory network of transcriptional regulators', *Nature*, vol. 403, no. 6767, pp. 335–338, Jan. 2000, doi: 10.1038/35002125.

[61] J. Schnakenberg, 'Simple chemical reaction systems with limit cycle behaviour', *J. Theor. Biol.*, vol. 81, no. 3, pp. 389–400, Dec. 1979, doi: 10.1016/0022-5193(79)90042-0.

[62] 'van der Pol Oscillator', in *Essentials of Mathematica*, New York, NY: Springer New York, 2007, pp. 505–508.

[63] O. Purcell, N. J. Savery, C. S. Grierson, and M. di Bernardo, 'A comparative analysis of synthetic genetic oscillators', *J. R. Soc. Interface*, vol. 7, no. 52, pp. 1503–1524, Nov. 2010, doi: 10.1098/rsif.2010.0183.

[64] M. Han and P. Yu, 'Hopf Bifurcation and Normal Form Computation', in *Normal Forms, Melnikov Functions and Bifurcations of Limit Cycles*, vol. 181, London: Springer London, 2012, pp. 7–58.

[65] Wikipedia, 'Hopf Bifurcation'. [Online]. Available: https://en.wikipedia.org/wiki/Hopf_bifurcation.

[66] H. G. Brachtendorf, R. Melville, P. Feldmann, S. Lampe, and R. Laur, 'Homotopy Method for Finding the Steady States of Oscillators', *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 33, no. 6, pp. 867–878, Jun. 2014, doi: 10.1109/TCAD.2014.2302637.

[67] P. Langfield, B. Krauskopf, and H. M. Osinga, 'Solving Winfree's puzzle: The isochrons in the FitzHugh-Nagumo model', *Chaos Interdiscip. J. Nonlinear Sci.*, vol. 24, no. 1, p. 013131, Mar. 2014, doi: 10.1063/1.4867877.

[68] C. Canavier, 'Phase response curve', *Scholarpedia*, vol. 1, no. 12, p. 1332, 2006, doi: 10.4249/scholarpedia.1332.

[69] M. H. Vitaterna *et al.*, 'The mouse Clock mutation reduces circadian pacemaker amplitude and enhances efficacy of resetting stimuli and phase-response curve amplitude', *Proc. Natl. Acad. Sci.*, vol. 103, no. 24, pp. 9327–9332, Jun. 2006, doi: 10.1073/pnas.0603601103.

[70] A. Pikovsky, M. Rosenblum, and J. Kurths, 'PHASE SYNCHRONIZATION IN REGULAR AND CHAOTIC SYSTEMS', *Int. J. Bifurc. Chaos*, vol. 10, no. 10, pp. 2291–2305, Oct. 2000, doi: 10.1142/S0218127400001481.

[71] M. Heltberg, R. A. Kellogg, S. Krishna, S. Tay, and M. H. Jensen, 'Noise Induces Hopping between NF-κB Entrainment Modes', *Cell Syst.*, vol. 3, no. 6, pp. 532-539.e3, Dec. 2016, doi: 10.1016/j.cels.2016.11.014.

[72] M. L. Heltberg and M. H. Jensen, 'Locked body clocks', *Nat. Phys.*, vol. 15, no. 10, pp. 989–990, Oct. 2019, doi: 10.1038/s41567-019-0617-2.

[73] M. G. Rosenblum and A. S. Pikovsky, 'Detecting direction of coupling in interacting oscillators.', *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 64, no. 4 Pt 2, p. 045202, Oct. 2001.

[74] J. A. Acebrón, L. L. Bonilla, C. J. Pérez Vicente, F. Ritort, and R. Spigler, 'The Kuramoto model: A simple paradigm for synchronization phenomena', *Rev. Mod. Phys.*, vol. 77, no. 1, pp. 137–185, Apr. 2005, doi: 10.1103/RevModPhys.77.137.

[75] M. Wacker and H. Witte, 'On the Stability of the n:m Phase Synchronization Index', *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 332–338, Feb. 2011, doi: 10.1109/TBME.2010.2063028.

[76] C. O. Diekman and D. B. Forger, 'Clustering Predicted by an Electrophysiological Model of the Suprachiasmatic Nucleus', *J. Biol. Rhythms*, vol. 24, no. 4, pp. 322–333, Aug. 2009, doi: 10.1177/0748730409337601.

[77] Levins, R., 'The strategy of model building in population biology. American scientist', *American scientist*, 1966.

[78] 'Underfitting vs. Overfitting'. [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html.

[79] E.-J. Wagenmakers and S. Farrell, 'AIC model selection using Akaike weights', *Psychon. Bull. Rev.*, vol. 11, no. 1, pp. 192–196, Feb. 2004, doi: 10.3758/BF03206482.

[80] K. P. Burnham and D. R. Anderson, Eds., 'Monte Carlo Insights and Extended Examples', in *Model Selection and Multimodel Inference*, New York, NY: Springer New York, 2004, pp. 206–266.

[81] M. Chidambaram, *Mathematical modelling and simulation in chemical engineering*. Cambridge ; New York: Cambridge University Press, 2018.

[82] C. C. Aggarwal, *Neural networks and deep learning: a textbook*. Cham: Springer, 2018.

[83] T. Bohlin, *Practical grey-box process identification: theory and applications*. London: Springer, 2006.

[84] J. Monod, 'The Growth of Bacterial Cultures', *Annu. Rev. Microbiol.*, vol. 3, no. 1, pp. 371–394, Oct. 1949, doi: 10.1146/annurev.mi.03.100149.002103.

[85] J. Chambers, 'Developing a rapid, scalable method of thermal characterisation for UK dwellings using smart meter data', 2017, doi: 10.13140/RG.2.2.26587.62244.

[86] A. A. Neath and J. E. Cavanaugh, 'The Bayesian information criterion: background, derivation, and applications: The Bayesian information criterion', *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 4, no. 2, pp. 199–203, Mar. 2012, doi: 10.1002/wics.199.

[87] J. Kuha, 'AIC and BIC: Comparisons of Assumptions and Performance', *Sociol. Methods Res.*, vol. 33, no. 2, pp. 188–229, Nov. 2004, doi: 10.1177/0049124103262065.

[88] E.-J. Wagenmakers, M. Lee, T. Lodewyckx, and G. J. Iverson, 'Bayesian Versus Frequentist Inference', in *Bayesian Evaluation of Informative Hypotheses*, H. Hoijtink, I. Klugkist, and P. A. Boelen, Eds. New York, NY: Springer New York, 2008, pp. 181–207.

[89] H. Yanagisawa, O. Kawamata, and K. Ueda, 'Modeling Emotions Associated With Novelty at Variable Uncertainty Levels: A Bayesian Approach', *Front. Comput. Neurosci.*, vol. 13, p. 2, Jan. 2019, doi: 10.3389/fncom.2019.00002.

[90] W. R. Gilks, 'Markov Chain Monte Carlo', in *Encyclopedia of Biostatistics*, P. Armitage and T. Colton, Eds. Chichester, UK: John Wiley & Sons, Ltd, 2005, p. b2a14021.

[91] D. Delsuc, Frédéric Emmanuel, 'Les méthodes probabilistes en phylogénie moléculaire : l'approche bayésiene.' 2004.

[92] L. K. Fonken and R. J. Nelson, 'The effects of light at night on circadian clocks and metabolism', *Endocr. Rev.*, vol. 35, no. 4, pp. 648–670, Aug. 2014, doi: 10.1210/er.2013-1051.

[93] S. Masri, M. Cervantes, and P. Sassone-Corsi, 'The circadian clock and cell cycle: interconnected biological circuits', *Curr. Opin. Cell Biol.*, vol. 25, no. 6, pp. 730–734, Dec. 2013, doi: 10.1016/j.ceb.2013.07.013.

[94] D. Sage, M. Unser, P. Salmon, and C. Dibner, 'A software solution for recording circadian oscillator features in time-lapse live cell microscopy', *Cell Div.*, vol. 5, p. 17, Jul. 2010, doi: 10.1186/1747-1028-5-17.

[95] A. Papagiannakis, B. Niebel, E. C. Wit, and M. Heinemann, 'Autonomous Metabolic Oscillations Robustly Gate the Early and Late Cell Cycle', *Mol. Cell*, vol. 65, no. 2, pp. 285–295, Jan. 2017, doi: 10.1016/j.molcel.2016.11.018.

[96] C. Feillet, G. T. J. van der Horst, F. Levi, D. A. Rand, and F. Delaunay, 'Coupling between the Circadian Clock and Cell Cycle Oscillators: Implication for Healthy Cells and Malignant Growth', *Front. Neurol.*, vol. 6, May 2015, doi: 10.3389/fneur.2015.00096.

[97] R. Y. Moore and V. B. Eichler, 'Loss of a circadian adrenal corticosterone rhythm following suprachiasmatic lesions in the rat', *Brain Res.*, vol. 42, no. 1, pp. 201–206, Jul. 1972, doi: 10.1016/0006-8993(72)90054-6.

[98] F. K. Stephan and I. Zucker, 'Circadian rhythms in drinking behavior and locomotor activity of rats are eliminated by hypothalamic lesions', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 69, no. 6, pp. 1583–1586, Jun. 1972, doi: 10.1073/pnas.69.6.1583.

[99] C. Dibner, U. Schibler, and U. Albrecht, 'The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks', *Annu. Rev. Physiol.*, vol. 72, no. 1, pp. 517–549, Mar. 2010, doi: 10.1146/annurev-physiol-021909-135821.

[100] E. Nagoshi, C. Saini, C. Bauer, T. Laroche, F. Naef, and U. Schibler, 'Circadian gene expression in individual fibroblasts: cell-autonomous and self-sustained oscillators pass time to daughter cells', *Cell*, vol. 119, no. 5, pp. 693–705, Nov. 2004, doi: 10.1016/j.cell.2004.11.015.

[101] D. K. Welsh, S.-H. Yoo, A. C. Liu, J. S. Takahashi, and S. A. Kay, 'Bioluminescence Imaging of Individual Fibroblasts Reveals Persistent, Independently Phased Circadian Rhythms of Clock Gene Expression', *Curr. Biol.*, vol. 14, no. 24, pp. 2289–2295, Dec. 2004, doi: 10.1016/j.cub.2004.11.057.

[102] Raven, P.H, G.B. Johnson, K. A. Mason, J. B. Losos, and S. R. Singer, 'How Cells Divide', *Biology*, 2011.

[103] D. O. Morgan, *The cell cycle: principles of control*. London : Sunderland, MA: Published by New Science Press in association with Oxford University Press ; Distributed inside North America by Sinauer Associates, Publishers, 2007.

[104] M. V. Blagosklonny and A. B. Pardee, 'The restriction point of the cell cycle', *Cell Cycle Georget. Tex*, vol. 1, no. 2, pp. 103–110, Apr. 2002.

[105] P. M. Nurse, 'NOBEL LECTURE: Cyclin Dependent Kinases and Cell Cycle Control', *Biosci. Rep.*, vol. 22, no. 5–6, pp. 487–499, Dec. 2002, doi: 10.1023/A:1022017701871.

[106] A. Chaix, A. Zarrinpar, and S. Panda, 'The circadian coordination of cell biology', *J. Cell Biol.*, vol. 215, no. 1, pp. 15–25, Oct. 2016, doi: 10.1083/jcb.201603076.

[107] C. Levi, 'Mitosis: Biochemical Pathways'. Student Reader, Jun. 04, 2009, [Online]. Available: https://studentreader.com/L8613/mitosis-biochemical-pathways/.

[108] T. Hunt and P. Sassone-Corsi, 'Riding Tandem: Circadian Clocks and the Cell Cycle', *Cell*, vol. 129, no. 3, pp. 461–464, May 2007, doi: 10.1016/j.cell.2007.04.015.

[109] A. Gréchez-Cassiau, B. Rayet, F. Guillaumond, M. Teboul, and F. Delaunay, 'The Circadian Clock Component BMAL1 Is a Critical Regulator of $p21^{WAF1/CIP1}$ Expression and

Hepatocyte Proliferation', *J. Biol. Chem.*, vol. 283, no. 8, pp. 4535–4542, Feb. 2008, doi: 10.1074/jbc.M705576200.

[110] E. Kowalska *et al.*, 'NONO couples the circadian clock to the cell cycle', *Proc. Natl. Acad. Sci.*, vol. 110, no. 5, pp. 1592–1599, Jan. 2013, doi: 10.1073/pnas.1213317110.

[111] T. Matsuo, 'Control Mechanism of the Circadian Clock for Timing of Cell Division in Vivo', *Science*, vol. 302, no. 5643, pp. 255–259, Oct. 2003, doi: 10.1126/science.1086271.

[112] X. Yang, P. A. Wood, and W. J. M. Hrushesky, 'Mammalian TIMELESS is required for ATM-dependent CHK2 activation and G2/M checkpoint control', *J. Biol. Chem.*, vol. 285, no. 5, pp. 3030–3034, Jan. 2010, doi: 10.1074/jbc.M109.050237.

[113] T.-H. Kang and S.-H. Leem, 'Modulation of ATR-mediated DNA damage checkpoint response by cryptochrome 1', *Nucleic Acids Res.*, vol. 42, no. 7, pp. 4427–4434, Apr. 2014, doi: 10.1093/nar/gku094.

[114] J. Gottesfeld, 'Mitotic repression of the transcriptional machinery', *Trends Biochem. Sci.*, vol. 22, no. 6, pp. 197–202, Jun. 1997, doi: 10.1016/S0968-0004(97)01045-1.

[115] M. Oklejewicz, E. Destici, F. Tamanini, R. A. Hut, R. Janssens, and G. T. J. van der Horst, 'Phase Resetting of the Mammalian Circadian Clock by DNA Damage', *Curr. Biol.*, vol. 18, no. 4, pp. 286–291, Feb. 2008, doi: 10.1016/j.cub.2008.01.047.

[116] T. Miki, T. Matsumoto, Z. Zhao, and C. C. Lee, 'p53 regulates Period2 expression and the circadian clock', *Nat. Commun.*, vol. 4, no. 1, p. 2444, Dec. 2013, doi: 10.1038/ncomms3444.

[117] T. Mori and C. H. Johnson, 'Circadian control of cell division in unicellular organisms', in *Progress in Cell Cycle Research*, L. Meijer, A. Jézéquel, and B. Ducommun, Eds. Boston, MA: Springer US, 2000, pp. 185–192.

[118] A. Bolige, S. Hagiwara, Y. Zhang, and K. Goto, 'Circadian G2 Arrest as Related to Circadian Gating of Cell Population Growth in Euglena', *Plant Cell Physiol.*, vol. 46, no. 6, pp. 931–936, Jun. 2005, doi: 10.1093/pcp/pci100.

[119] M. Yeom, J. S. Pendergast, Y. Ohmiya, and S. Yamazaki, 'Circadian-independent cell mitosis in immortalized fibroblasts', *Proc. Natl. Acad. Sci.*, vol. 107, no. 21, pp. 9665–9670, May 2010, doi: 10.1073/pnas.0914078107.

[120] G. A. Bjarnason *et al.*, 'Circadian Expression of Clock Genes in Human Oral Mucosa and Skin', *Am. J. Pathol.*, vol. 158, no. 5, pp. 1793–1801, May 2001, doi: 10.1016/S0002-9440(10)64135-1.

[121] T. G. Granda *et al.*, 'Circadian regulation of cell cycle and apoptosis proteins in mouse bone marrow and tumor', *FASEB J.*, vol. 19, no. 2, pp. 1–22, Feb. 2005, doi: 10.1096/fj.04-2665fje.

[122] Takako Noguchi and Susan S. Golden, 'Bioluminescent and fluorescent reporters in circadian rhythm', 2017.

[123] A. Sakaue-Sawano *et al.*, 'Visualizing Spatiotemporal Dynamics of Multicellular Cell-Cycle Progression', *Cell*, vol. 132, no. 3, pp. 487–498, Feb. 2008, doi: 10.1016/j.cell.2007.12.033.

[124] O. Sandler, S. P. Mizrahi, N. Weiss, O. Agam, I. Simon, and N. Q. Balaban, 'Lineage correlations of single cell division time as a probe of cell-cycle dynamics', *Nature*, vol. 519, no. 7544, pp. 468–471, Mar. 2015, doi: 10.1038/nature14318.

[125] A. P. Dempster and N. M. Laird and D. B. Rubin, 'Maximum likelihood from incomplete data via the EM algorithm', 1977.

[126] J. A. Bilmes and others, 'A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models', *Int. Comput. Sci. Inst.*, vol. 4, no. 510, p. 126, 1998.

[127] C. F. J. Wu, 'On the Convergence Properties of the EM Algorithm', *Ann. Stat.*, vol. 11, no. 1, pp. 95–103, Mar. 1983, doi: 10.1214/aos/1176346060.

[128] Green, PJ, 'On use of the EM algorithm for penalized likelihood estimation', Jan. 1990.

[129] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*. New York ; London: Springer, 2005.

[130] Stamp, Mark, 'A revealing introduction to hidden markov models', Jan. 01, 2004.

[131] J. Mermet, J. Yeung, and F. Naef, 'Systems Chronobiology: Global Analysis of Gene Regulation in a 24-Hour Periodic World', *Cold Spring Harb. Perspect. Biol.*, vol. 9, no. 3, p. a028720, Jan. 2017, doi: 10.1101/cshperspect.a028720.

[132] A. Hahn, J. Jones, and T. Meyer, 'Quantitative analysis of cell cycle phase durations and PC12 differentiation using fluorescent biosensors.', *Cell Cycle*, vol. 8, pp. 1044–52, Apr. 2009.

[133] S. L. Spencer, S. D. Cappell, F.-C. Tsai, K. W. Overton, C. L. Wang, and T. Meyer, 'The proliferation-quiescence decision is controlled by a bifurcation in CDK2 activity at mitotic exit', *Cell*, vol. 155, no. 2, pp. 369–383, Oct. 2013, doi: 10.1016/j.cell.2013.08.062.

[134] T. Mori, B. Binder, and C. Johnson, 'Circadian gating of cell division in cyanobacteria growing with average doubling times of less than 24 hours.', *Proc Natl Acad Sci U A*, vol. 93, pp. 10183–8, Sep. 1996.

[135] Q. Yang, B. F. Pando, G. Dong, S. S. Golden, and A. van Oudenaarden, 'Circadian gating of the cell cycle revealed in single cyanobacterial cells.', *Science*, vol. 327, no. 5972, pp. 1522–6, Mar. 2010, doi: 10.1126/science.1181759.

[136] T. Matsu-Ura *et al.*, 'Intercellular Coupling of the Cell Cycle and Circadian Clock in Adult Stem Cell Culture', *Mol. Cell*, vol. 64, no. 5, pp. 900–912, 01 2016, doi: 10.1016/j.molcel.2016.10.015.

[137] M. V. Plikus *et al.*, 'Local circadian clock gates cell cycle progression of transient amplifying cells during regenerative hair cycling', *Proc. Natl. Acad. Sci.*, vol. 110, no. 23, pp. E2106–E2115, Jun. 2013, doi: 10.1073/pnas.1215935110.

[138] C. Gérard and A. Goldbeter, 'Entrainment of the mammalian cell cycle by the circadian clock: modeling two coupled cellular rhythms', *PLoS Comput. Biol.*, vol. 8, no. 5, p. e1002516, May 2012, doi: 10.1371/journal.pcbi.1002516.

[139] J. Paijmans, M. Bosman, P. R. ten Wolde, and D. K. Lubensky, 'Discrete gene replication events drive coupling between the cell cycle and circadian clocks', *Proc. Natl. Acad. Sci.*, vol. 113, no. 15, pp. 4063–4068, Apr. 2016, doi: 10.1073/pnas.1507291113.

[140] J. Rougemont and F. Naef, 'Collective synchronization in populations of globally coupled phase oscillators with drifting frequencies', *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 73, no. 1 Pt 1, p. 011104, Jan. 2006, doi: 10.1103/PhysRevE.73.011104.

[141] L. Rabiner and B. Juang, 'An introduction to hidden Markov models', *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, 1986, doi: 10.1109/MASSP.1986.1165342.

[142] C. Vollmers, S. Panda, and L. DiTacchio, 'A High-Throughput Assay for siRNA-Based Circadian Screens in Human U2OS Cells', *PLoS ONE*, vol. 3, no. 10, p. e3457, Oct. 2008, doi: 10.1371/journal.pone.0003457.

[143] B. Maier *et al.*, 'A large-scale functional RNAi screen reveals a role for CK2 in the mammalian circadian clock', *Genes Dev.*, vol. 23, no. 6, pp. 708–718, Mar. 2009, doi: 10.1101/gad.512209.

[144] D. Nicolas, B. Zoller, D. M. Suter, and F. Naef, 'Modulation of transcriptional burst frequency by histone acetylation', *Proc. Natl. Acad. Sci.*, vol. 115, no. 27, pp. 7153–7158, Jul. 2018, doi: 10.1073/pnas.1722330115.

[145] A. Balsalobre, 'Resetting of Circadian Time in Peripheral Tissues by Glucocorticoid Signaling', *Science*, vol. 289, no. 5488, pp. 2344–2347, Sep. 2000, doi: 10.1126/science.289.5488.2344.

[146] C. Saini, J. Morf, M. Stratmann, P. Gos, and U. Schibler, 'Simulated body temperature rhythms reveal the phase-shifting behavior and plasticity of mammalian circadian oscillators', *Genes Dev.*, vol. 26, no. 6, pp. 567–580, Mar. 2012, doi: 10.1101/gad.183251.111.

[147] O. Hayes, B. Ramos, L. L. Rodríguez, A. Aguilar, T. Badía, and F. O. Castro, 'Cell confluency is as efficient as serum starvation for inducing arrest in the G0/G1 phase of the cell cycle in granulosa and fibroblast cells of cattle', *Anim. Reprod. Sci.*, vol. 87, no. 3–4, pp. 181–192, Jul. 2005, doi: 10.1016/j.anireprosci.2004.11.011.

[148] J. Yeung *et al.*, 'Transcription factor activity rhythms and tissue-specific chromatin interactions explain circadian gene expression across organs', *Genome Res.*, vol. 28, no. 2, pp. 182–191, Feb. 2018, doi: 10.1101/gr.222430.117.

[149] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, 'A circadian gene expression atlas in mammals: implications for biology and medicine', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 45, pp. 16219–16224, Nov. 2014, doi: 10.1073/pnas.1408886111.

[150] N. Cermakian, 'Altered behavioral rhythms and clock gene expression in mice with a targeted mutation in the Period1 gene', *EMBO J.*, vol. 20, no. 15, pp. 3967–3974, Aug. 2001, doi: 10.1093/emboj/20.15.3967.

[151] J. P. DeBruyne, E. Noton, C. M. Lambert, E. S. Maywood, D. R. Weaver, and S. M. Reppert, 'A Clock Shock: Mouse CLOCK Is Not Required for Circadian Oscillator Function', *Neuron*, vol. 50, no. 3, pp. 465–477, May 2006, doi: 10.1016/j.neuron.2006.03.041.

[152] A. E. Granada and H. Herzel, 'How to Achieve Fast Entrainment? The Timescale to Synchronization', *PLoS ONE*, vol. 4, no. 9, p. e7057, Sep. 2009, doi: 10.1371/journal.pone.0007057.

[153] R. P. Aryal *et al.*, 'Macromolecular Assemblies of the Mammalian Circadian Clock', *Mol. Cell*, vol. 67, no. 5, pp. 770-782.e6, Sep. 2017, doi: 10.1016/j.molcel.2017.07.017.

[154] M. Ukai-Tadenuma, R. G. Yamada, H. Xu, J. A. Ripperger, A. C. Liu, and H. R. Ueda, 'Delay in Feedback Repression by Cryptochrome 1 Is Required for Circadian Clock Function', *Cell*, vol. 144, no. 2, pp. 268–281, Jan. 2011, doi: 10.1016/j.cell.2010.12.019.

[155] J. Mermet *et al.*, 'Clock-dependent chromatin topology modulates circadian transcription and behavior', *Genes Dev.*, vol. 32, no. 5–6, pp. 347–358, Mar. 2018, doi: 10.1101/gad.312397.118.

[156] S. M. Siepka *et al.*, 'Circadian Mutant Overtime Reveals F-box Protein FBXL3 Regulation of Cryptochrome and Period Gene Expression', *Cell*, vol. 129, no. 5, pp. 1011–1023, Jun. 2007, doi: 10.1016/j.cell.2007.04.030.

[157] R. W. King, R. J. Deshaies, J.-M. Peters, and M. W. Kirschner, 'How Proteolysis Drives the Cell Cycle', *Science*, vol. 274, no. 5293, pp. 1652–1659, Dec. 1996, doi: 10.1126/science.274.5293.1652.

[158] M. H. Hastings, E. S. Maywood, and M. Brancaccio, 'Generation of circadian rhythms in the suprachiasmatic nucleus', *Nat. Rev. Neurosci.*, vol. 19, no. 8, pp. 453–469, Aug. 2018, doi: 10.1038/s41583-018-0026-z.

[159] J. A. Mohawk, C. B. Green, and J. S. Takahashi, 'Central and Peripheral Circadian Clocks in Mammals', *Annu. Rev. Neurosci.*, vol. 35, no. 1, pp. 445–462, Jul. 2012, doi: 10.1146/annurev-neuro-060909-153128.

[160] J. Shilts, G. Chen, and J. J. Hughey, 'Evidence for widespread dysregulation of circadian clock progression in human cancer', *PeerJ*, vol. 6, p. e4327, Jan. 2018, doi: 10.7717/peerj.4327.

[161] F. C. Kelleher, A. Rao, and A. Maguire, 'Circadian molecular clocks and cancer', *Cancer Lett.*, vol. 342, no. 1, pp. 9–18, Jan. 2014, doi: 10.1016/j.canlet.2013.09.040.

[162] S. L. Brunton, J. L. Proctor, and J. N. Kutz, 'Discovering governing equations from data by sparse identification of nonlinear dynamical systems', *Proc. Natl. Acad. Sci.*, vol. 113, no. 15, pp. 3932–3937, Apr. 2016, doi: 10.1073/pnas.1517384113.

[163] D. Soroldoni *et al.*, 'A Doppler effect in embryonic pattern formation', *Science*, vol. 345, no. 6193, pp. 222–225, Jul. 2014, doi: 10.1126/science.1253089.

[164] M. I. Koksharov and N. N. Ugarova, 'APPROACHES TO ENGINEER STABILITY OF BEETLE LUCIFERASES', *Comput. Struct. Biotechnol. J.*, vol. 2, no. 3, p. e201204004, Sep. 2012, doi: 10.5936/csbj.201209004.

[165] D. W. Huang, B. T. Sherman, and R. A. Lempicki, 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009, doi: 10.1093/nar/gkn923.

[166] D. S. Lemons and P. Langevin, *An introduction to stochastic processes in physics: containing 'On the theory of Brownian motion' by Paul Langevin, translated by Anthony Gythiel*. Baltimore: Johns Hopkins University Press, 2002.

[167] K. Norrman, Y. Fischer, B. Bonnamy, F. Wolfhagen Sand, P. Ravassard, and H. Semb, 'Quantitative Comparison of Constitutive Promoters in Human ES cells', *PLoS ONE*, vol. 5, no. 8, p. e12413, Aug. 2010, doi: 10.1371/journal.pone.0012413.

[168] J. Y. Qin *et al.*, 'Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter', *PLoS ONE*, vol. 5, no. 5, p. e10611, May 2010, doi: 10.1371/journal.pone.0010611.

[169] D. Mauvoisin *et al.*, 'Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver', *Proc. Natl. Acad. Sci.*, vol. 111, no. 1, pp. 167–172, Jan. 2014, doi: 10.1073/pnas.1314066111.

[170] K. Bahar Halpern *et al.*, 'Bursty gene expression in the intact mammalian liver', *Mol. Cell*, vol. 58, no. 1, pp. 147–156, Apr. 2015, doi: 10.1016/j.molcel.2015.01.027.

[171] Y. Lin, S. Nosaka, Y. Amakata, and T. Maeda, 'Comparative study of the mammalian liver innervation: an immunohistochemical study of protein gene product 9.5, dopamine β-hydroxylase and tyrosine hydroxylase', *Comp. Biochem. Physiol. A Physiol.*, vol. 110, no. 4, pp. 289–298, Apr. 1995, doi: 10.1016/0300-9629(94)00189-Z.

[172] T. Katsuda *et al.*, 'Transcriptomic Dissection of Hepatocyte Heterogeneity: Linking Ploidy, Zonation, and Stem/Progenitor Cell Characteristics', *Cell. Mol. Gastroenterol. Hepatol.*, vol. 9, no. 1, pp. 161–183, 2020, doi: 10.1016/j.jcmgh.2019.08.011.

[173] K. B. Halpern *et al.*, 'Single-cell spatial reconstruction reveals global division of labour in the mammalian liver', *Nature*, vol. 542, no. 7641, pp. 352–356, 16 2017, doi: 10.1038/nature21065.

[174] F. Atger *et al.*, 'Circadian and feeding rhythms differentially affect rhythmic mRNA transcription and translation in mouse liver', *Proc. Natl. Acad. Sci.*, vol. 112, no. 47, pp. E6579–E6588, Nov. 2015, doi: 10.1073/pnas.1515308112.

[175] P. Godoy *et al.*, 'Recent advances in 2D and 3D in vitro systems using primary hepatocytes, alternative hepatocyte sources and non-parenchymal liver cells and their use in investigating mechanisms of hepatotoxicity, cell signaling and ADME', *Arch. Toxicol.*, vol. 87, no. 8, pp. 1315–1530, Aug. 2013, doi: 10.1007/s00204-013-1078-5.

[176] R. Gebhardt and D. Mecke, 'Heterogeneous distribution of glutamine synthetase among rat liver parenchymal cells in situ and in primary culture', *EMBO J.*, vol. 2, no. 4, pp. 567–570, 1983.

[177] D. Sasse, N. Katz, and K. Jungermann, 'Functional heterogeneity of rat liver parenchyma and of isolated hepatocytes', *FEBS Lett.*, vol. 57, no. 1, pp. 83–88, Sep. 1975.

[178] R. Gebhardt, 'Metabolic zonation of the liver: Regulation and implications for liver function', *Pharmacol. Ther.*, vol. 53, no. 3, pp. 275–354, Jan. 1992, doi: 10.1016/0163-7258(92)90055-5.

[179] K. Jungermann and T. Kietzmann, 'Zonation of parenchymal and nonparenchymal metabolism in liver', *Annu. Rev. Nutr.*, vol. 16, pp. 179–203, 1996, doi: 10.1146/annurev.nu.16.070196.001143.

[180] S. Ben-Moshe and S. Itzkovitz, 'Spatial heterogeneity in the mammalian liver', *Nat. Rev. Gastroenterol. Hepatol.*, vol. 16, no. 7, pp. 395–410, Jul. 2019, doi: 10.1038/s41575-019-0134-x.

[181] G. Zajicek, R. Oren, and M. Weinreb, 'The streaming liver', *Liver*, vol. 5, no. 6, pp. 293–300, Dec. 1985.

[182] M. P. Bralet, S. Branchereau, C. Brechot, and N. Ferry, 'Cell lineage study in the liver using retroviral mediated gene transfer. Evidence against the streaming of hepatocytes in normal liver', *Am. J. Pathol.*, vol. 144, no. 5, pp. 896–905, May 1994.

[183] T. Kietzmann, 'Metabolic zonation of the liver: The oxygen gradient revisited', *Redox Biol.*, vol. 11, pp. 622–630, 2017, doi: 10.1016/j.redox.2017.01.012.

[184] K. Jungermann and T. Kietzmann, 'Oxygen: Modulator of metabolic zonation and disease of the liver', *Hepatology*, vol. 31, no. 2, pp. 255–260, Feb. 2000, doi: 10.1002/hep.510310201.

[185] A. B. Novikoff, 'SYMPOSIUM: The Biochemical Cytology of Liver', *J. Histochem. Cytochem.*, vol. 7, no. 4, pp. 213–213, Jul. 1959, doi: 10.1177/7.4.213.

[186] D. L. Schmucker, J. S. Mooney, and A. L. Jones, 'Stereological analysis of hepatic fine structure in the Fischer 344 rat. Influence of sublobular location and animal age', *J. Cell Biol.*, vol. 78, no. 2, pp. 319–337, Aug. 1978.

[187] T. Kietzmann, Y. Cornesse, K. Brechtel, S. Modaressi, and K. Jungermann, 'Perivenous expression of the mRNA of the three hypoxia-inducible factor alpha-subunits, HIF1alpha, HIF2alpha and HIF3alpha, in rat liver', *Biochem. J.*, vol. 354, no. Pt 3, pp. 531–537, Mar. 2001.

[188] S. T. Koury, M. C. Bondurant, M. J. Koury, and G. L. Semenza, 'Localization of cells producing erythropoietin in murine liver by in situ hybridization', *Blood*, vol. 77, no. 11, pp. 2497–2503, Jun. 1991.

[189] S. Benhamouche *et al.*, 'Apc tumor suppressor gene is the "zonation-keeper" of mouse liver', *Dev. Cell*, vol. 10, no. 6, pp. 759–770, Jun. 2006, doi: 10.1016/j.devcel.2006.03.015.

[190] G. Zeng *et al.*, 'Wnt'er in liver: expression of Wnt and frizzled genes in mouse', *Hepatol. Baltim. Md*, vol. 45, no. 1, pp. 195–204, Jan. 2007, doi: 10.1002/hep.21473.

[191] K. Matsumoto, R. Miki, M. Nakayama, N. Tatsumi, and Y. Yokouchi, 'Wnt9a secreted from the walls of hepatic sinusoids is essential for morphogenesis, proliferation, and glycogen accumulation of chick hepatic epithelium', *Dev. Biol.*, vol. 319, no. 2, pp. 234–247, Jul. 2008, doi: 10.1016/j.ydbio.2008.04.021.

[192] A. S. Rocha *et al.*, 'The Angiocrine Factor Rspondin3 Is a Key Determinant of Liver Zonation', *Cell Rep.*, vol. 13, no. 9, pp. 1757–1764, Dec. 2015, doi: 10.1016/j.celrep.2015.10.049.

[193] S. Neumann *et al.*, 'Mammalian Wnt3a is released on lipoprotein particles', *Traffic Cph. Den.*, vol. 10, no. 3, pp. 334–343, Mar. 2009, doi: 10.1111/j.1600-0854.2008.00872.x.

[194] M. Matz-Soja, A. Hovhannisyan, and R. Gebhardt, 'Hedgehog signalling pathway in adult liver: a major new player in hepatocyte metabolism and zonation?', *Med. Hypotheses*, vol. 80, no. 5, pp. 589–594, May 2013, doi: 10.1016/j.mehy.2013.01.032.

[195] A. D. Güler *et al.*, 'Melanopsin cells are the principal conduits for rod-cone input to non-image-forming vision', *Nature*, vol. 453, no. 7191, pp. 102–105, May 2008, doi: 10.1038/nature06829.

[196] I. Provencio, G. Jiang, W. J. De Grip, W. P. Hayes, and M. D. Rollag, 'Melanopsin: An opsin in melanophores, brain, and eye', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 1, pp. 340–345, Jan. 1998.

[197] R. Greger and M. Bleich, 'Normal Values for Physiological Parameters', in *Comprehensive Human Physiology: From Cellular Mechanisms to Integration*, R. Greger and U. Windhorst, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 2427–2449.

[198] M. R. Ralph, R. G. Foster, F. C. Davis, and M. Menaker, 'Transplanted suprachiasmatic nucleus determines circadian period', *Science*, vol. 247, no. 4945, pp. 975–978, Feb. 1990.

[199] F. Damiola, N. Le Minh, N. Preitner, B. Kornmann, F. Fleury-Olela, and U. Schibler, 'Restricted feeding uncouples circadian oscillators in peripheral tissues from the central pacemaker in the suprachiasmatic nucleus', *Genes Dev.*, vol. 14, no. 23, pp. 2950–2961, Dec. 2000.

[200] K. A. Stokkan, S. Yamazaki, H. Tei, Y. Sakaki, and M. Menaker, 'Entrainment of the circadian clock in the liver by feeding', *Science*, vol. 291, no. 5503, pp. 490–493, Jan. 2001, doi: 10.1126/science.291.5503.490.

[201] Y. Adamovich, B. Ladeuix, M. Golik, M. P. Koeners, and G. Asher, 'Rhythmic Oxygen Levels Reset Circadian Clocks through HIF1α', *Cell Metab.*, vol. 25, no. 1, pp. 93–101, Jan. 2017, doi: 10.1016/j.cmet.2016.09.014.

[202] I. N. Karatsoreos and R. Silver, 'Body Clocks in Health and Disease', in *Conn's Translational Neuroscience*, Elsevier, 2017, pp. 599–615.

[203] S. Panda *et al.*, 'Coordinated transcription of key pathways in the mouse by the circadian clock', *Cell*, vol. 109, no. 3, pp. 307–320, May 2002.

[204] C. Vollmers, S. Gill, L. DiTacchio, S. R. Pulivarthy, H. D. Le, and S. Panda, 'Time of feeding and the intrinsic circadian clock drive rhythms in hepatic gene expression', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 50, pp. 21453–21458, Dec. 2009, doi: 10.1073/pnas.0909591106.

[205] D. Feng *et al.*, 'A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism', *Science*, vol. 331, no. 6022, pp. 1315–1319, Mar. 2011, doi: 10.1126/science.1198125.

[206] T. Li and J. Y. L. Chiang, 'Bile Acid Signaling in Metabolic Disease and Drug Therapy', *Pharmacol. Rev.*, vol. 66, no. 4, pp. 948–983, Oct. 2014, doi: 10.1124/pr.113.008201.

[207] N. R. Katz, W. Fischer, and S. Giffhorn, 'Distribution of enzymes of fatty acid and ketone body metabolism in periportal and perivenous rat-liver tissue', *Eur. J. Biochem.*, vol. 135, no. 1, pp. 103–107, Sep. 1983.

[208] J. L. Evans, B. Quistorff, and L. A. Witters, 'Zonation of hepatic lipogenic enzymes identified by dual-digitonin-pulse perfusion', *Biochem. J.*, vol. 259, no. 3, pp. 821–829, May 1989.

[209] M. Guzmán and J. Castro, 'Zonation of fatty acid metabolism in rat liver', *Biochem. J.*, vol. 264, no. 1, pp. 107–113, Nov. 1989, doi: 10.1042/bj2640107.

[210] J. Schleicher *et al.*, 'Zonation of hepatic fatty acid metabolism — The diversity of its regulation and the benefit of modeling', *Biochim. Biophys. Acta BBA - Mol. Cell Biol. Lipids*, vol. 1851, no. 5, pp. 641–656, May 2015, doi: 10.1016/j.bbalip.2015.02.004.

[211] Jean-Louis Foulley, 'Mixed Model Methodology, Part I: Linear Mixed Models', 2015, doi: 10.13140/2.1.3072.0320.

[212] Michael Freeman, 'A Visual Introduction to Hierarchical Modeling'. 2017, [Online]. Available: https://github.com/mkfreeman/hierarchical-models/.

[213] S. Müller, J. L. Scealy, and A. H. Welsh, 'Model Selection in Linear Mixed Models', *Stat. Sci.*, vol. 28, no. 2, pp. 135–167, May 2013, doi: 10.1214/12-STS410.

[214] R. Gebhardt, 'Metabolic zonation of the liver: regulation and implications for liver function', *Pharmacol. Ther.*, vol. 53, no. 3, pp. 275–354, 1992.

[215] S. Hoehme *et al.*, 'Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration', *Proc. Natl. Acad. Sci.*, vol. 107, no. 23, pp. 10371–10376, Jun. 2010, doi: 10.1073/pnas.0909374107.

[216] B. Wang, L. Zhao, M. Fish, C. Y. Logan, and R. Nusse, 'Self-renewing diploid Axin2+ cells fuel homeostatic renewal of the liver', *Nature*, vol. 524, no. 7564, pp. 180–185, Aug. 2015, doi: 10.1038/nature14863.

[217] L. Planas-Paz *et al.*, 'The RSPO–LGR4/5–ZNRF3/RNF43 module controls liver zonation and size', *Nat. Cell Biol.*, vol. 18, no. 5, pp. 467–479, May 2016, doi: 10.1038/ncb3337.

[218] C. Dibner, U. Schibler, and U. Albrecht, 'The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks', *Annu. Rev. Physiol.*, vol. 72, no. 1, pp. 517–549, 2010, doi: 10.1146/annurev-physiol-021909-135821.

[219] C. B. Green, J. S. Takahashi, and J. Bass, 'The Meter of Metabolism', *Cell*, vol. 134, no. 5, pp. 728–742, Sep. 2008, doi: 10.1016/j.cell.2008.08.022.

[220] J. Yeung and F. Naef, 'Rhythms of the Genome: Circadian Dynamics from Chromatin Topology, Tissue-Specific Gene Expression, to Behavior', *Trends Genet. TIG*, vol. 34, no. 12, pp. 915–926, Dec. 2018, doi: 10.1016/j.tig.2018.09.005.

[221] E. Maury, K. M. Ramsey, and J. Bass, 'Circadian Rhythms and Metabolic Syndrome: From Experimental Genetics to Human Disease', *Circ. Res.*, vol. 106, no. 3, pp. 447–462, Feb. 2010, doi: 10.1161/CIRCRESAHA.109.208355.

[222] J. A. Sobel *et al.*, 'Transcriptional regulatory logic of the diurnal cycle in the mouse liver', *PLOS Biol.*, vol. 15, no. 4, p. e2001069, Apr. 2017, doi: 10.1371/journal.pbio.2001069.

[223] J. Wang *et al.*, 'Nuclear Proteomics Uncovers Diurnal Regulatory Landscapes in Mouse Liver', *Cell Metab.*, vol. 25, no. 1, pp. 102–117, Jan. 2017, doi: 10.1016/j.cmet.2016.10.003.

[224] K. B. Halpern *et al.*, 'Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells', *Nat. Biotechnol.*, vol. 36, no. 10, pp. 962–970, Nov. 2018, doi: 10.1038/nbt.4231.

[225] J. Beal, 'Biochemical complexity drives log-normal variation in genetic expression', *Eng. Biol.*, vol. 1, no. 1, pp. 55–60, Jun. 2017, doi: 10.1049/enb.2017.0004.

[226] R. A. McLean, W. L. Sanders, and W. W. Stroup, 'A Unified Approach to Mixed Linear Models', *Am. Stat.*, vol. 45, no. 1, p. 54, Feb. 1991, doi: 10.2307/2685241.

[227] F. Atger *et al.*, 'Circadian and feeding rhythms differentially affect rhythmic mRNA transcription and translation in mouse liver', *Proc. Natl. Acad. Sci.*, vol. 112, no. 47, pp. E6579–E6588, Nov. 2015, doi: 10.1073/pnas.1515308112.

[228] K. Jungermann, 'Dynamics of zonal hepatocyte heterogeneity. Perinatal development and adaptive alterations during regeneration after partial hepatectomy, starvation and diabetes', *Acta Histochem. Suppl.*, vol. 32, pp. 89–98, 1986.

[229] D. W. Huang, B. T. Sherman, and R. A. Lempicki, 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, Jan. 2009, doi: 10.1038/nprot.2008.211.

[230] R. Gebhardt, 'Liver zonation: Novel aspects of its regulation and its impact on homeostasis', *World J. Gastroenterol.*, vol. 20, no. 26, p. 8491, 2014, doi: 10.3748/wjg.v20.i26.8491.

[231] N. Aizarani *et al.*, 'A human liver cell atlas reveals heterogeneity and epithelial progenitors', *Nature*, vol. 572, no. 7768, pp. 199–204, Aug. 2019, doi: 10.1038/s41586-019-1373-2.

[232] J. Mermet *et al.*, 'Clock-dependent chromatin topology modulates circadian transcription and behavior', *Genes Dev.*, vol. 32, no. 5–6, pp. 347–358, Mar. 2018, doi: 10.1101/gad.312397.118.

[233] G. Le Martelot *et al.*, 'Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles', *PLoS Biol.*, vol. 10, no. 11, p. e1001442, Nov. 2012, doi: 10.1371/journal.pbio.1001442.

[234] J. Wang *et al.*, 'Circadian clock-dependent and -independent posttranscriptional regulation underlies temporal mRNA accumulation in mouse liver', *Proc. Natl. Acad. Sci.*, vol. 115, no. 8, pp. E1916–E1925, Feb. 2018, doi: 10.1073/pnas.1715225115.

[235] M. Nitzan, N. Karaiskos, N. Friedman, and N. Rajewsky, 'Gene expression cartography', *Nature*, Nov. 2019, doi: 10.1038/s41586-019-1773-3.

[236] P. O. Seglen, 'Chapter 4 Preparation of Isolated Rat Liver Cells', in *Methods in Cell Biology*, vol. 13, Elsevier, 1976, pp. 29–83.

[237] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, 'Spatial reconstruction of single-cell gene expression data', *Nat. Biotechnol.*, vol. 33, no. 5, pp. 495–502, May 2015, doi: 10.1038/nbt.3192.

[238] A. V. Loud, 'A QUANTITATIVE STEREOLOGICAL DESCRIPTION OF THE ULTRASTRUCTURE OF NORMAL RAT LIVER PARENCHYMAL CELLS', *J. Cell Biol.*, vol. 37, no. 1, pp. 27–46, Apr. 1968, doi: 10.1083/jcb.37.1.27.

[239] N. J. Kuhn, M. Woodworth-Gutai, K. W. Gross, and W. A. Held, 'Subfamilies of the mouse major urinary protein (MUP) multi-gene family: sequence analysis of cDNA clones and differential regulation in the liver', *Nucleic Acids Res.*, vol. 12, no. 15, pp. 6073–6090, 1984, doi: 10.1093/nar/12.15.6073.

[240] S. Müller, J. L. Scealy, and A. H. Welsh, 'Model Selection in Linear Mixed Models', *Stat. Sci.*, vol. 28, no. 2, pp. 135–167, May 2013, doi: 10.1214/12-STS410.

[241] G. Schwarz, 'Estimating the Dimension of a Model', *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978, doi: 10.1214/aos/1176344136.

[242] S. R. Jammalamadaka and A. SenGupta, *Topics in Circular Statistics*, vol. 5. WORLD SCIENTIFIC, 2001.

[243] A. Lyubimova, S. Itzkovitz, J. P. Junker, Z. P. Fan, X. Wu, and A. van Oudenaarden, 'Single-molecule mRNA detection and counting in mammalian tissue', *Nat. Protoc.*, vol. 8, no. 9, pp. 1743–1758, Aug. 2013, doi: 10.1038/nprot.2013.109.

[244] S. Purewal, *Learning web app development*, First edition. Beijing: O'Reilly, 2014.

[245] Plotly team, 'Python Dash, web-based analytic apps with Python'. 2020, [Online]. Available: https://plotly.com/dash/.

[246] RStudio, Inc, 'ShinyR: Easy web application in R.' [Online]. Available: http://www.rstudio.com/shiny/.

[247] Y. Wang and N. E. Navin, 'Advances and Applications of Single-Cell Sequencing Technologies', *Mol. Cell*, vol. 58, no. 4, pp. 598–609, May 2015, doi: 10.1016/j.molcel.2015.05.005.

[248] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann, 'The Technology and Biology of Single-Cell RNA Sequencing', *Mol. Cell*, vol. 58, no. 4, pp. 610–620, May 2015, doi: 10.1016/j.molcel.2015.04.005.

[249] R. Cannoodt, W. Saelens, and Y. Saeys, 'Computational methods for trajectory inference from single-cell transcriptomics', *Eur. J. Immunol.*, vol. 46, no. 11, pp. 2496–2506, Nov. 2016, doi: 10.1002/eji.201646347.

[250] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein, 'Fundamental limits on dynamic inference from single-cell snapshots', *Proc. Natl. Acad. Sci.*, vol. 115, no. 10, pp. E2467–E2476, Mar. 2018, doi: 10.1073/pnas.1714723115.

[251] G. L. Manno *et al.*, 'RNA velocity of single cells', *Nature*, p. 1, Aug. 2018, doi: 10.1038/s41586-018-0414-6.

[252] A. R. Lederer and G. La Manno, 'The emergence and promise of single-cell temporal-omics approaches', *Curr. Opin. Biotechnol.*, vol. 63, pp. 70–78, Jun. 2020, doi: 10.1016/j.copbio.2019.12.005.

[253] A. Kulkarni, A. G. Anderson, D. P. Merullo, and G. Konopka, 'Beyond bulk: a review of single cell transcriptomics methodologies and applications', *Curr. Opin. Biotechnol.*, vol. 58, pp. 129–136, 2019, doi: 10.1016/j.copbio.2019.03.001.

[254] N. Aizarani *et al.*, 'A human liver cell atlas reveals heterogeneity and epithelial progenitors', *Nature*, vol. 572, no. 7768, pp. 199–204, Aug. 2019, doi: 10.1038/s41586-019-1373-2.

[255] Tabula Muris Consortium *et al.*, 'Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris', *Nature*, vol. 562, no. 7727, pp. 367–372, 2018, doi: 10.1038/s41586-018-0590-4.

[256] Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium *et al.*, 'Genetic identification of brain cell types underlying schizophrenia', *Nat. Genet.*, vol. 50, no. 6, pp. 825–833, Jun. 2018, doi: 10.1038/s41588-018-0129-5.

[257] N. Moris, C. Pina, and A. M. Arias, 'Transition states and cell fate decisions in epigenetic landscapes', *Nat. Rev. Genet.*, vol. 17, no. 11, pp. 693–703, Nov. 2016, doi: 10.1038/nrg.2016.98.

[258] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, 'Exponential scaling of single-cell RNA-seq in the past decade', *Nat. Protoc.*, vol. 13, no. 4, pp. 599–604, Apr. 2018, doi: 10.1038/nprot.2017.149.

[259] A. Tanay and A. Regev, 'Scaling single-cell genomics from phenomenology to mechanism', *Nature*, vol. 541, no. 7637, pp. 331–338, Jan. 2017, doi: 10.1038/nature21350.

[260] J. Packer and C. Trapnell, 'Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation', *Trends Genet.*, vol. 34, no. 9, pp. 653–665, Sep. 2018, doi: 10.1016/j.tig.2018.06.001.

[261] G.-J. Hendriks *et al.*, 'NASC-seq monitors RNA synthesis in single cells', *Nat. Commun.*, vol. 10, no. 1, p. 3138, Dec. 2019, doi: 10.1038/s41467-019-11028-9.

[262] A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure, 'Whole-organism lineage tracing by combinatorial and cumulative genome editing', *Science*, vol. 353, no. 6298, p. aaf7907, Jul. 2016, doi: 10.1126/science.aaf7907.

[263] K. L. Frieda *et al.*, 'Synthetic recording and in situ readout of lineage information in single cells', *Nature*, vol. 541, no. 7635, pp. 107–111, Jan. 2017, doi: 10.1038/nature20777.

[264] O. Guillaume-Gentil *et al.*, 'Tunable Single-Cell Extraction for Molecular Analyses', *Cell*, vol. 166, no. 2, pp. 506–516, Jul. 2016, doi: 10.1016/j.cell.2016.06.025.

[265] C. Trapnell *et al.*, 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nat. Biotechnol.*, vol. 32, no. 4, pp. 381–386, Apr. 2014, doi: 10.1038/nbt.2859.

[266] S. C. Bendall *et al.*, 'Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development', *Cell*, vol. 157, no. 3, pp. 714–725, Apr. 2014, doi: 10.1016/j.cell.2014.04.005.

[267] S. Liang, F. Wang, J. Han, and K. Chen, 'Latent periodic process inference from single-cell RNA-seq data', *Nat. Commun.*, vol. 11, no. 1, p. 1441, Dec. 2020, doi: 10.1038/s41467-020-15295-9.

[268] S. Huang, 'The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?', *BioEssays*, vol. 34, no. 2, pp. 149–157, Feb. 2012, doi: 10.1002/bies.201100031.

[269] V. Svensson and L. Pachter, 'RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq', *Mol. Cell*, vol. 72, no. 1, pp. 7–9, Oct. 2018, doi: 10.1016/j.molcel.2018.09.026.

[270] N. Battich *et al.*, 'Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies', *Science*, vol. 367, no. 6482, pp. 1151–1156, Mar. 2020, doi: 10.1126/science.aax3072.

[271] Q. Lo Giudice, M. Leleu, G. La Manno, and P. J. Fabre, 'Single-cell transcriptional logic of cell-fate specification and axon guidance in early-born retinal neurons', *Development*, vol. 146, no. 17, p. dev178103, Sep. 2019, doi: 10.1242/dev.178103.

[272] S. Kanton *et al.*, 'Organoid single-cell genomic atlas uncovers human-specific features of brain development', *Nature*, vol. 574, no. 7778, pp. 418–422, Oct. 2019, doi: 10.1038/s41586-019-1654-9.

[273] Q. Zhang *et al.*, 'Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma', *Cell*, vol. 179, no. 4, pp. 829-845.e20, Oct. 2019, doi: 10.1016/j.cell.2019.10.003.

[274] J. C. Kimmel, L. Penland, N. D. Rubinstein, D. G. Hendrickson, D. R. Kelley, and A. Z. Rosenthal, 'A murine aging cell atlas reveals cell identity and tissue-specific trajectories of aging', Bioinformatics, preprint, Jun. 2019. doi: 10.1101/657726.

[275] A. M. Klein, 'Technique to measure the expression dynamics of each gene in a single cell', *Nature*, vol. 560, no. 7719, pp. 434–435, Aug. 2018, doi: 10.1038/d41586-018-05882-8.

[276] E. K. P. Chong and S. H. Żak, *An introduction to optimization*, Fourth edition. Hoboken, New Jersey: Wiley, 2013.

[277] J. L. Morales, 'A numerical study of limited memory BFGS methods', *Appl. Math. Lett.*, vol. 15, no. 4, pp. 481–487, May 2002, doi: 10.1016/S0893-9659(01)00162-8.

[278] I. Kanter and T. Kalisky, 'Single Cell Transcriptomics: Methods and Applications', *Front. Oncol.*, vol. 5, Mar. 2015, doi: 10.3389/fonc.2015.00053.

[279] J. A. Griffiths, A. Scialdone, and J. C. Marioni, 'Using single-cell genomics to understand developmental processes and cell fate decisions', *Mol. Syst. Biol.*, vol. 14, no. 4, p. e8046, 16 2018, doi: 10.15252/msb.20178046.

[280] D. Lähnemann *et al.*, 'Eleven grand challenges in single-cell data science', *Genome Biol.*, vol. 21, no. 1, p. 31, Dec. 2020, doi: 10.1186/s13059-020-1926-6.

[281] A. McDavid, G. Finak, and R. Gottardo, 'The contribution of cell cycle to heterogeneity in single-cell RNA-seq data', *Nat. Biotechnol.*, vol. 34, no. 6, pp. 591–593, Jun. 2016, doi: 10.1038/nbt.3498.

[282] C. P. Fall, Ed., *Computational cell biology*. New York: Springer, 2002.

[283] A. Bastidas-Ponce *et al.*, 'Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis', *Development*, vol. 146, no. 12, Jun. 2019, doi: 10.1242/dev.173849.

[284] F. A. Wolf, P. Angerer, and F. J. Theis, 'SCANPY: large-scale single-cell gene expression data analysis', *Genome Biol.*, vol. 19, no. 1, p. 15, Dec. 2018, doi: 10.1186/s13059-017-1382-0.

[285] E. Z. Macosko *et al.*, 'Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets', *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015, doi: 10.1016/j.cell.2015.05.002.

[286] H. Mizuno, Y. Nakanishi, N. Ishii, A. Sarai, and K. Kitada, 'A signature-based method for indexing cell cycle phase distribution from microarray profiles', *BMC Genomics*, vol. 10, no. 1, p. 137, 2009, doi: 10.1186/1471-2164-10-137.

[287] L. Y. Chan, C. F. Mugler, S. Heinrich, P. Vallotton, and K. Weis, 'Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability', *eLife*, vol. 7, p. e32536, Sep. 2018, doi: 10.7554/eLife.32536.

[288] T. A. Geddes *et al.*, 'Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis', *BMC Bioinformatics*, vol. 20, no. S19, p. 660, Dec. 2019, doi: 10.1186/s12859-019-3179-5.

[289] E. Lin, S. Mukherjee, and S. Kannan, 'A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis', *BMC Bioinformatics*, vol. 21, no. 1, p. 64, Dec. 2020, doi: 10.1186/s12859-020-3401-5.

[290] V. Svensson, A. Gayoso, N. Yosef, and L. Pachter, 'Interpretable factor models of single-cell RNA-seq via variational autoencoders', *Bioinformatics*, vol. 36, no. 11, pp. 3418–3421, Jun. 2020, doi: 10.1093/bioinformatics/btaa169.

[291] T. J. McGrew, M. Alspector-Kelly, and F. Allhoff, Eds., *The philosophy of science: an historical anthology*. Chichester, U.K. ; Malden, MA: Wiley-Blackwell, 2009.

[292] E. J. Olsson, 'Bayesian Epistemology', in *Introduction to Formal Philosophy*, S. O. Hansson and V. F. Hendricks, Eds. Cham: Springer International Publishing, 2018, pp. 431–442.

[293] B. Russell, *The problems of philosophy*. New York: Oxford University Press, 1997.

[294] D. P. Feldman, *Chaos and dynamical systems*. 2019.

[295] L. Chen, R.-S. Wang, and X.-S. Zhang, *Biomolecular networks: methods and applications in systems biology*. Hoboken, N.J: Wiley, 2009.

[296] L. H. Nguyen and S. Holmes, 'Ten quick tips for effective dimensionality reduction', *PLOS Comput. Biol.*, vol. 15, no. 6, p. e1006907, Jun. 2019, doi: 10.1371/journal.pcbi.1006907.

[297] B.-K. Liao, D. J. Jörg, and A. C. Oates, 'Faster embryonic segmentation through elevated Delta-Notch signalling', *Nat. Commun.*, vol. 7, no. 1, p. 11861, Sep. 2016, doi: 10.1038/ncomms11861.

[298] P. Congdon, *Bayesian hierarchical models: with applications using R*, Second edition. Boca Raton: CRC Press, 2020.

[299] S. Webb, 'Deep learning for biology', *Nature*, vol. 554, no. 7693, pp. 555–557, Feb. 2018, doi: 10.1038/d41586-018-02174-z.

[300] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, 'Deep learning microscopy', *Optica*, vol. 4, no. 11, p. 1437, Nov. 2017, doi: 10.1364/OPTICA.4.001437.

[301] GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, Granada, Spain *et al.*, 'Deep Learning in Omics Data Analysis and Precision Medicine', in *Computational Biology*, Division of Biomedical Science, University of the Highlands and Islands, UK and H. Husi, Eds. Codon Publications, 2019, pp. 37–53.

[302] G. Montavon, W. Samek, and K.-R. Müller, 'Methods for interpreting and understanding deep neural networks', *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018, doi: 10.1016/j.dsp.2017.10.011.

[303] R. Cestnik and M. Rosenblum, 'Inferring the phase response curve from observation of a continuously perturbed oscillator', *Sci. Rep.*, vol. 8, no. 1, p. 13606, Dec. 2018, doi: 10.1038/s41598-018-32069-y.

[304] R. Dahlhaus, T. Dumont, S. Le Corff, and J. C. Neddermeyer, 'Statistical inference for oscillation processes', *Statistics*, vol. 51, no. 1, pp. 61–83, Jan. 2017, doi: 10.1080/02331888.2016.1266985.

[305] Y. Luo, S. Al-Dossary, M. Marhoon, and M. Alfaraj, 'Generalized Hilbert transform and its applications in geophysics', *Lead. Edge*, vol. 22, no. 3, pp. 198–202, Mar. 2003, doi: 10.1190/1.1564522.

[306] W. M. Moon, A. Ushah, V. Singh, and B. Bruce, 'Application of 2-D Hilbert transform in geophysical imaging with potential field data', *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 5, pp. 502–510, Sep. 1988, doi: 10.1109/36.7674.

[307] W. Freeman, 'Hilbert transform for brain waves', *Scholarpedia*, vol. 2, no. 1, p. 1338, 2007, doi: 10.4249/scholarpedia.1338.

[308] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge, U.K. ; New York: Cambridge University Press, 2013.

[309] A. Zeiler, R. Faltermeier, I. R. Keck, A. M. Tome, C. G. Puntonet, and E. W. Lang, 'Empirical Mode Decomposition - an introduction', in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona, Spain, Jul. 2010, pp. 1–8, doi: 10.1109/IJCNN.2010.5596829.

[310] A. Boggess and F. J. Narcowich, *A first course in wavelets with Fourier analysis*, 2nd ed. Hoboken, N.J: John Wiley & Sons, 2009.

[311] N. Zielke and B. A. Edgar, 'FUCCI sensors: powerful new tools for analysis of cell proliferation: FUCCI sensors', *Wiley Interdiscip. Rev. Dev. Biol.*, vol. 4, no. 5, pp. 469–487, Sep. 2015, doi: 10.1002/wdev.189.

[312] R. F. Galán, G. B. Ermentrout, and N. N. Urban, 'Efficient estimation of phase-resetting curves in real neurons and its significance for neural-network modeling', *Phys. Rev. Lett.*, vol. 94, no. 15, p. 158101, Apr. 2005, doi: 10.1103/PhysRevLett.94.158101.

[313] T. Imai, K. Ota, and T. Aoyagi, 'Robust Measurements of Phase Response Curves Realized via Multicycle Weighted Spike-Triggered Averages', *J. Phys. Soc. Jpn.*, vol. 86, no. 2, p. 024009, Feb. 2017, doi: 10.7566/JPSJ.86.024009.

[314] G. Zhang, Y. Li, T. Wang, H. Du, F. Luo, and Y. Zhan, 'Extended Hilbert Transform and Application for Seismic Attributes', *Earth Space Sci.*, p. 2019EA000551, May 2019, doi: 10.1029/2019EA000551.

[315] M. Azmani, S. Reboul, J.-B. Choquel, and M. Benjelloun, 'A recursive fusion filter for angular data', in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Guilin, China, Dec. 2009, pp. 882–887, doi: 10.1109/ROBIO.2009.5420492.

[316] Chen Muyi and Wang Hongyuan, 'A new recursive filter based on the Gauss von Mises distribution', in *2015 IEEE 6th International Symposium on Microwave, Antenna, Propagation, and EMC Technologies (MAPE)*, Shanghai, China, Oct. 2015, pp. 329–332, doi: 10.1109/MAPE.2015.7510325.

[317] G. Kurz, I. Gilitschenski, and U. D. Hanebeck, 'Recursive nonlinear filtering for angular data based on circular distributions', in *2013 American Control Conference*, Washington, DC, Jun. 2013, pp. 5439–5445, doi: 10.1109/ACC.2013.6580688.

[318] M. P. Deisenroth and H. Ohlsson, 'A general perspective on Gaussian filtering and smoothing: Explaining current and deriving new algorithms', in *Proceedings of the 2011 American Control Conference*, San Francisco, CA, Jun. 2011, pp. 1807–1812, doi: 10.1109/ACC.2011.5990871.

[319] Roger R. Labbe Jr, 'Kalman and Bayesian Filters in Python'. 2015, [Online]. Available: https://github.com/rlabbe/Kalman-and-Bayesian-Filters-in-Python.

[320] C. Droin *et al.*, 'Space-time logic of liver gene expression at sublobular scale', Systems Biology, preprint, Mar. 2020. doi: 10.1101/2020.03.05.976571.

Curriculum Vitae

Colas Droin

PhD student in mathematical biology

Route de la Pierre 5
1024 Ecublens
Switzerland
✉ *colas.droin@epfl.ch*
27 years old – Driving license
Nationality : French

Education

October 2016 – Present
PhD in computational systems biology lab, *EPFL*, Lausanne, Switzerland.
- Subject: Single-cell analysis of noisy biological oscillators and their interactions.
- Supervisor: Pr Felix Naef, Computational Systems Biology laboratory.
- Relevant courses: Deep Learning, Symmetry and Conservation in the Cell, Responsible Conduct in Biomedical Research.

September 2015 – June 2016
Master of Science in *Complex Systems*, *Ecole Normale Supérieure (**ENS**) de Lyon*, France.
- **Graduated with honors**
- Relevant courses: complex networks, statistical physics, dynamical processes and networks, network algorithms for molecular biology, rule-based modeling.

September 2014 – January 2015
Exchange student in the school of Computer Science of the *Royal Institute of Technology*, *KTH*, Stockholm, Sweden.
- Relevant courses: discrete mathematics, artificial intelligence, Markov models, population genetics.

September 2012 – June 2015
Master degree in *Computational Biology and Mathematics*, *Institut National des Sciences Appliquées (**INSA**) de Lyon*, France.
- **Graduated with honors**
- Relevant courses: software development, ODE, PDE, statistics, cellular biology.

September 2010 – June 2012
Preparatory class, *INSA de Lyon*, France.
- Intensive two-year undergraduate courses in engineering sciences

Research schools

April 2019
CompSysBio, ***Aussois***, France, 1 week.
Methodological issues in computational systems biology.

August 2017
Synthetic & Systems Biology Summer School, ***Cambridge***, UK, 1 week.
Advances and methods in synthetic biology and systems biology.

January 2016
Statistical Modeling and Learning from Data, ***ENS***, Lyon, 1 week.
Applied machine learning.

December 2015
Randomized Algorithms, ***ENS***, Lyon, 1 week.
Decision problems, stochastic processes, theory of learning.

Internships

February 2016 – June 2016
ENS master thesis in the team Dante, *IXXI-ENS*, Lyon, 4.5 months.
- Subject: Data-driven characterization of social tie heterogeneities in real information cascades.
- Details: Exploration and extraction of data in a corpus containing several billions of tweets, social network construction and analysis (SQLite, Python, R).

February 2015 – July 2015
INSA master thesis in the team Biovia, *Dassault Systèmes*, Vélizy, 6 months.
- Subject: study and development of a hybrid system for the Chemical Master Equation.
- Details: state of the art of hybrid systems (differential-stochastic), choice of an optimal model, integration of a boolean feature, development, optimization and tests (C++, R).

May 2014 – August 2014
Intern in the team robotics of the Stem Cell and Brain Research Institute, *INSERM u846*, Lyon, France, 3 months.
- Subject: study of a sentences production model based on a recurrent neural network.
- Details: model analysis, debugging and optimization (Python).

July 2013 – August 2013 **Intern in the mixed recherch unit MISTEA**, *INRA-INRIA complex*, Montpellier, France, 5 weeks.
- Subject: mathematical and numerical modeling of a soil ecosystem
- Details: model development (Python), Equation analysis and numerical simulations

Teaching and tutoring

Fall 2018-2019-2020 **Computational and mathematical modelling in biology**, *Bachelor level*, Teaching assistant, EPFL, Lausanne, Switzerland.
- Head TA: management and scheduling of the TA sessions
- Exercise redesign: conversion of the exercises into interactive Jupyter notebooks
- Project (homework) conception and correction

Spring 2017 **Probabilities and statistics II.**, *Bachelor level*, Teaching assistant, EPFL, Lausanne, Switzerland.

October 2012 – June 2014 **General tutoring**, *Bachelor level*, INSA Passerelle Program, Lyon, France.
Personalized assistance to students with learning difficulties

Talks

20/03/2020 **RNA velocity-based inference of cell cycle properties using single-cells**, Physics of living systems, EPFL, Lausanne, Switzerland.

12/12/2019 **Understanding the space-time logic of the mammalian liver with mixed-models**, SSS, EPFL, Lausanne, Switzerland.

17/06/2019 **Using HMM to understand the dynamics of two coupled oscillators**, Functional Genomics, EPFL, Lausanne, Switzerland.

03/04/2019 **Low-dimensional Dynamics of Two Coupled Biological Oscillators**, CompSysBio, CNRS, Aussois, France.

20/07/2017 **A reconstruction of the phase dynamics of interacting cell-cycle and circadian clock**, SSBSS, Cambridge, UK.

16/06/2017 **Characterization of the influence of the cell-cycle in the circadian clock in mammalian cells**, Physics of living systems, EPFL, Lausanne, Switzerland.

Posters

26/11/2019 **Space-time logic of the mammalian liver**, Theoretical Biology Network in Western Switzerland, UNIL, Lausanne, Switzerland..

10/04/2019 **Exploring circadian liver zonation**, DDay, CHUV, Lausanne, Switzerland..

14/03/2019 **Exploring circadian liver zonation**, VISualizing BIological data 2019 (VIZBI). Heidelberg, Germany..

01/10/2018 **Deciphering the Phase Dynamics of Two Coupled Biological Oscillators**, EPFL Physics days, Lausanne, Switzerland..

04/06/2018 **Low-dimensional Dynamics of Two Coupled Biological Oscillators**, EMBO/ EMBL Symposium on Biological Oscillators: Design, Mechanism, Function. Heidelberg, Germany..

04/08/2017 **How can two biological oscillators synchronize? A study of the cell-cycle influence on the circadian clock**, 3rd International SystemsX.ch Conference on Systems Biology, Zurich, Switzerland..

20/07/2017 **A reconstruction of the phase dynamics of interacting cell-cycle and circadian clock**, SSBSS, Cambridge, UK.

Publications

2020 **Space-time logic of liver gene expression at sublobular scale**, *C. Droin, J. El Kholtei, K. B. Halpen, F. Naef, S. Itzkovitz*, Currently being reviewed.

2019 **Low-dimensional Dynamics of Two Coupled Biological Oscillators**, *C. Droin, E. Paquet, F. Naef*, Nature Physics.

2015 **Corticostriatal response selection in sentence production: Insights from neural network simulation with reservoir computing**, *Hinaut, X., Lance, F., Droin, C., Petit, M., Pointeau, G., & Dominey, P. F.*, 2015, Brain and language.

Mainstream media

Summer 2019 **Coupling project mentioned in several popular journals and on the radio, including *Le Matin* and *RTS1***.

Awards

10/04/2019 **Best poster award**, DDay, CHUV, Lausanne, Switzerland.

01/10/2018 **Best poster award**, EPFL Physics days. Lausanne, Switzerland.

Computer science skills

OS	Proficient with OS-X and Windows, good knowledge of Linux
Programming	Python, R, Julia, C, C++
Web	HTML5, CSS, Bootstrap, Python Dash, Processing, D3
Graphism	Adobe Illustrator, Adobe Photoshop
Office	Office suite (Word, Powerpoint, Excel) and equivalents, Keynote, Latex, Beamer

Languages

French	Native tongue
English	Fluent
Spanish	Intermediate

External work and hobbies

September 2017-Present **Data science blogging**, *http://colasdroin.eu/DataScienceBlog/*.
Various topics regarding bio-physics, machine learning and statistics

September 2018 **Release of my first book**, *Bittersweet*, IS Editions.

September 2011 – June 2013 **Communication manager**, *INSA badminton's association*, Lyon, France.
Creation of posters, management of a newsletter, events organization

References

Available upon request.